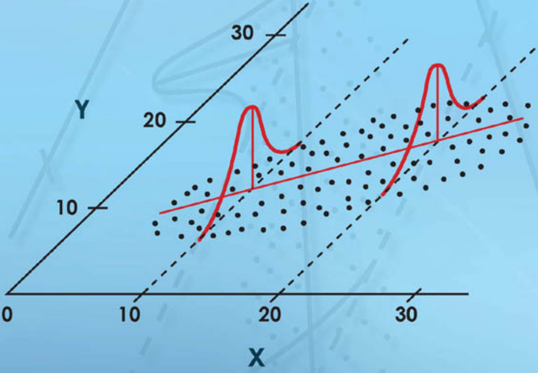
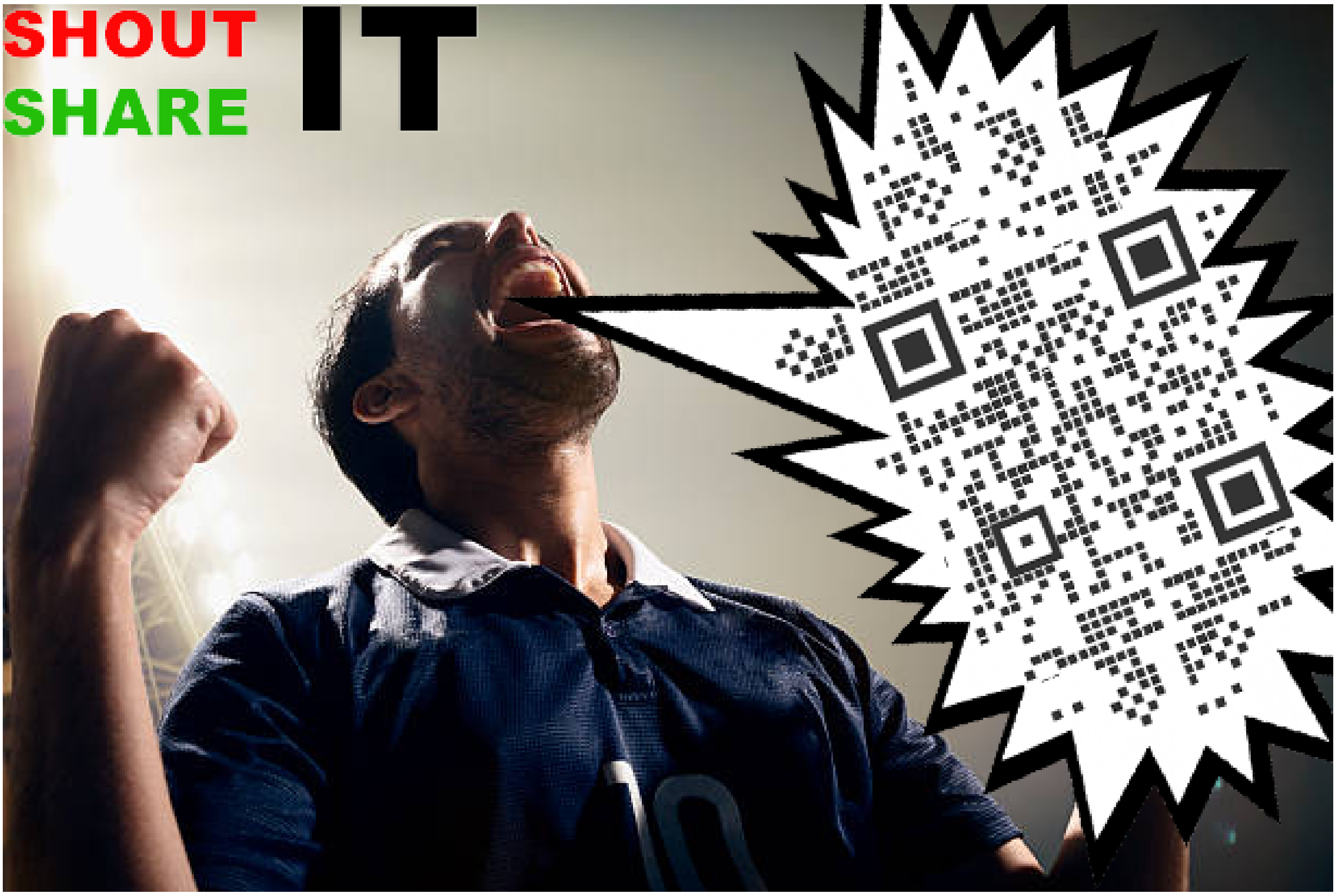
SIXTH
EDITION

Statistical Reasoning in the Behavioral Sciences

BRUCE M. KING • PATRICK J. ROSOPA • EDWARD W. MINIUM


30
Y
20
10
0
10
20
30
X


SHOUT
SHARE
IT

This page intentionally left blank

Statistical Reasoning in the Behavioral Sciences

This page intentionally left blank

Statistical Reasoning in the Behavioral Sciences

Sixth Edition

Bruce M. King
Patrick J. Rosopa
Edward W. Minium

WILEY
John Wiley & Sons, Inc.

Vice President and Publisher	Jay O'Callaghan
Editor	Robert Johnston
Associate Editor	Eileen McKeever
Editorial Assistant	Mariah Maguire-Fong
Marketing Manager	Danielle Torio
Designer	RDC Publishing Group Sdn Bhd
Production Manager	Janis Soo
Senior Production Editor	Joyce Poh

This book was set in 10/12 Bembo by Thomson Digital and printed and bound by RR Donnelley.
The cover was printed by RR Donnelley.

This book is printed on acid free paper.

Copyright © 2011, 2008, 2003, 1993 John Wiley & Sons, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc. 222 Rosewood Drive, Danvers, MA 01923, website www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, (201)748-6011, fax (201)748-6008, website http://www.wiley.com/go/permissions.

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at www.wiley.com/go/returnlabel. Outside of the United States, please contact your local representative.

Library of Congress Cataloging-in-Publication Data
King, Bruce M.
Statistical reasoning in the behavioral sciences / Bruce M. King, Patrick Rosopa, Edward W. Minium. — 6th ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-64382-2 (hardback)
1. Psychometrics. 2. Educational statistics. I. Rosopa, Patrick. II. Minium, Edward W., 1917- III. Title.
BF39.K53 2010
300.72'7—dc22
2010020656

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

To Students Who Want to Learn
and Instructors Who Like to Teach

This page intentionally left blank

PREFACE

Statistical Reasoning in the Behavioral Sciences is emphatically concerned with students understanding of statistical logic and procedures, as well as an understanding of what happens when the strict requirements of statistical theory meet the circumstances of real-world data. Regardless of one's academic field, it is important that students understand the interplay between statistical questions and answers and research questions and answers so that students see that statistics really has significance for their discipline. Thus, throughout the text we interweave aspects of experimental design with the statistical topics to which they are closely related. We try to be especially clear on those points in inferential statistics that tend to be stumbling blocks. At the same time, we give full treatment in the first half of the text to descriptive statistics in the belief that once inference is done, one must return to the descriptive measures to assess the meaning of the inquiry; both concepts must be fully understood.

The fourth and fifth editions of *Statistical Reasoning* followed the recommendations of the APA Task Force on Statistical Inference (Wilkinson et al., 1999) and we continue to do so in the sixth edition. This includes greater emphasis on effect size, power, and confidence intervals. Beginning with Chapter 13, measures of effect size (both *d* family and *r* family measures) are integrated with the material on hypothesis testing.

The text should serve very well for a one-semester course. Even so, there is more material than can be covered in most one-quarter or one-semester courses. The Instructor's Manual offers suggestions for adapting the book to particular needs.

Features New to the Sixth Edition

In many introductory courses, instructors require that students learn how to use computers to work statistical problems. Thus, the major change from previous editions is the addition of instructions for the use of the statistical software package IBM SPSS. Step-by-step instructions are provided as well as 25 screen shots so that students may easily use the software. We have also made an attempt to update many of the example and end-of-chapter problems in the textbook. The only organizational change was the placement of the chapter on chi-square after the chapters on ANOVA in order not to interrupt the presentation of statistical methods used with quantitative variables (from the single-sample t test through ANOVA).

Although we demonstrate how to use a powerful, user-friendly software package, it is important to highlight that we do not sacrifice any of the material dealing with the conceptual understanding of a statistical procedure. For example, after thoroughly describing the logic behind a procedure, we complete a problem step-by-step using raw data and formulas, followed by an interpretation of the results. Only after adequate conceptual coverage and a complete example do we demonstrate how to use IBM SPSS (i.e., entering data, navigating menus, and using dialog boxes) to obtain the same results in seconds. For instructors who prefer not to cover IBM SPSS, these sections can be skipped without disrupting the continuity of a chapter.

Pedagogy

Statistical Reasoning in the Behavioral Sciences includes the following pedagogical features:

- learning objectives at the beginning of each chapter,
- a summary at the end of each chapter to reinforce what has been presented,
- key terms and concepts that are boldfaced at their first occurrence in each chapter and listed again for review at the end of each chapter,
- the presentation of definitions, symbols, and equations in the margins of the text so that the student may easily identify important new material,
- special boxed sections called Points of Controversy that introduce students to topical and/or current issues to let them see that statistics is an evolving, vital area of inquiry, and
- review of basic mathematical propositions in an appendix.

Some texts present chapters on correlation and regression after the chapters on hypothesis testing. However, we firmly believe that the material on correlation should come first so that we may more fully develop the material on dependent-groups designs and the r family of measures of effect size.

Finally, many students who enroll in introductory statistics do so with great anxiety. Many statistics textbooks lack any visual material other than graphs of frequency distributions and mathematical tables and formulas. We include a few carefully chosen cartoons with hopes that they will make using the book more enjoyable and less frightening to students.

Homework and Test Problems/Instructor's Manual

There are numerous problems at the end of each chapter. Some give practice in problem-solving technique; others require the use of critical judgment. Answers to the odd-numbered problems are in an appendix at the end of the text. Answers to the even-numbered problems are in the Instructor's Manual so that instructors may assign the problems as homework or in-class exercises. The Instructor's Manual also contains over 700 additional questions. They include several hundred multiple-choice questions for quizzes and several data sets in each chapter appropriate for examinations.

The Instructor's Manual is available to instructors at www.wiley.com/college/king.

Acknowledgments

Many reviewers, mentors, colleagues, friends, and family members helped develop the text. To them we extend our lasting gratitude for debts that can never be fully repaid. In particular, we thank Ed Minium, who was sole author on this book when it was first published in 1970. His

vision and understanding of statistics have guided us ever since. Bruce King thanks his wife, Gail, for her support during the seemingly never-ending time they juggled the demands of authorship, teaching, and research. Patrick Rosopa would like to thank his wife, Melodee, for her love and support, bearing with him while he managed the demands of authorship, research, and teaching. He would also like to acknowledge his daughters, Elenah and Cecilia, whose love serves as a constant source of strength. We are especially grateful for the leadership of Eileen McKeever, whose confidence and patience enabled us to complete the project.

In addition, we want to express our gratitude to the following reviewers for suggestions made during various stages of the manuscript. We may not have made every change the reviewers desired, but each suggestion caused us to stop and carefully reassess our goals and the needs of the students and instructors. The reviewers are:

Robert W. Allan, Lafayette College
Jeffrey Andre, James Madison University
Razia Azen, University of Illinois at Urbana–Champaign
Frank M. Bagrash, California State University–Fullerton
Dr. D. K. Beale, Cerrito College
Dr. David M. Boynton, Western Connecticut State University
Shirley Brown, SUNY Albany
Jerry Cohen, University of Rhode Island
Jane Conner, SUNY Binghamton
Stephen S. Cooper
Donna Cruse
Ernest C. Davenport, University of Minnesota–Twin Cities
Jessica Dennis, California State University, Los Angeles
Janis W. Driscoll, University of Colorado at Denver
Bruce Dudek, SUNY Albany
Rick M. Gardner, University of Colorado at Denver
Sonia V. Gonsalves, Richard Stockton College of NJ
Sue A. Joslyn, University of Northern Iowa
Shelia Kennison, Oklahoma State University
Dr. Melanie Kercher, Sam Houston State University
Donald L. Meyer, University of Pittsburgh
Frederick L. Oswald, Michigan State University
Robert R. Provine, University of Maryland–Baltimore County
Robert M. Pruzek, SUNY Albany
William P. Reich, Loyola University of Chicago
Alan Reifman, Texas Tech University
William M. Rogers, Grand Valley State University
Margaret G. Ruddy, The College of New Jersey
Ron Serlin, University of Wisconsin–Madison
Murray Singer, University of Manitoba
Gregory T. Smith, University of Kentucky
Alan Sockloff, Temple University

John E. Stecklein, University of Minnesota
Dr. Chehalis Strapp, Western Oregon University
Lawrence J. Stricker, The New School for Social Research
Ellen P. Susman, Metropolitan State College
Penelope G. Vinden, Clark University
Julia Wallace, University of Northern Iowa
William Yost, Loyola University of Chicago
Dr. Lynnette Zellezny, California State University at Fresno

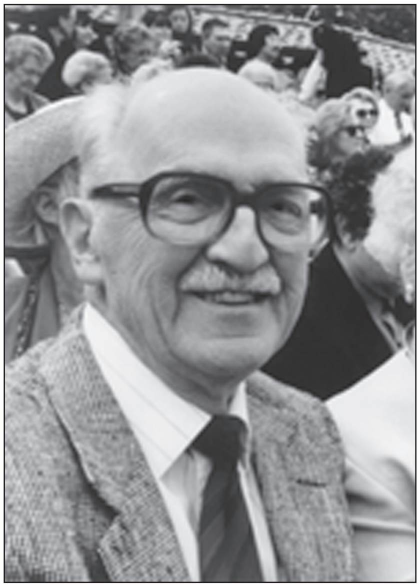
About the Authors

Bruce M. King 1946–

Bruce King earned his B.A. in psychology at UCLA in 1969 and a Ph.D. in biopsychology at the University of Chicago in 1978. He was a faculty member at the University of New Orleans from 1978 to 2007 and is presently at Clemson University. He has been honored as a Fellow in the Association for Psychological Science, American Psychological Association, and International Behavioral Neuroscience Society.

Patrick J. Rosopa 1972–

Patrick Rosopa earned his B.S. (2000) in psychology at Tulane University and his M.S. (2003) and Ph.D. (2006) in industrial/organizational psychology at the University of Central Florida. He is presently at Clemson University. The results of his research have been presented at refereed conferences held by the Academy of Management, American Psychological Association, Psychometric Society, and Society for Industrial and Organizational Psychology. He has published in such outlets as the *Encyclopedia of Measurement and Statistics*, *Journal of Vocational Behavior*, *Organizational Research Methods*, and *Personality and Individual Differences*.



Edward W. Minium 1917–2004

Ed Minium earned his B.A. at Stanford University in 1939. After serving with the United States Air Force during World War II, he returned to earn his Ph.D. in psychology at the University of California at Berkeley in 1951. He was a faculty member at San José State University from 1948 to 1982 and served as chair of the Department of Psychology from 1961 to 1966. He was the only author for the first two editions of this textbook, which was first published in 1970.

This page intentionally left blank

CONTENTS

CHAPTER 1
Introduction 1

1.1 Descriptive Statistics 3
1.2 Inferential Statistics 3
1.3 Our Concern: Applied Statistics 4
1.4 Variables and Constants 6
1.5 Scales of Measurement 7
1.6 Scales of Measurement and Problems of Statistical Treatment 9
1.7 Do Statistics Lie? 10
Point of Controversy *Are Statistical Procedures Necessary?* 11
1.8 Some Tips on Studying Statistics 12
1.9 Statistics and Computers 13
1.10 Summary 14

CHAPTER 2
Frequency Distributions, Percentiles, and Percentile Ranks 17
2.1 Organizing Qualitative Data 18
2.2 Grouped Scores 19
2.3 How to Construct a Grouped Frequency Distribution 21
2.4 Apparent versus Real Limits 22
2.5 The Relative Frequency Distribution 22
2.6 The Cumulative Frequency Distribution 24

2.7 Percentiles and Percentile Ranks 25
2.8 Computing Percentiles from Grouped Data 26
2.9 Computation of Percentile Rank 28
2.10 Summary 29

CHAPTER 3
Graphic Representation of Frequency Distributions 33
3.1 Basic Procedures 34
3.2 The Histogram 34
3.3 The Frequency Polygon 36
3.4 Choosing between a Histogram and a Polygon 37
3.5 The Bar Diagram and the Pie Chart 38
3.6 The Cumulative Percentage Curve 41
3.7 Factors Affecting the Shape of Graphs 42
3.8 Shape of Frequency Distributions 43
3.9 Summary 45

CHAPTER 4
Central Tendency 48
4.1 The Mode 49
4.2 The Median 49
4.3 The Mean 50
4.4 Properties of the Mode 51
4.5 Properties of the Mean 52
Point of Controversy *Is It Permissible to Calculate the Mean for Tests in the Behavioral Sciences?* 53
4.6 Properties of the Median 55
4.7 Measures of Central Tendency in Symmetrical and Asymmetrical Distributions 56
4.8 The Effects of Score Transformations 57
4.9 Summary 57

CHAPTER 5
Variability and Standard (z) Scores 61
5.1 The Range and Semi-Interquartile Range 63
5.2 Deviation Scores 64
5.3 Deviational Measures: The Variance 64
5.4 Deviational Measures: The Standard Deviation 66
5.5 Calculation of the Variance and Standard Deviation: Raw-Score Method 66
5.6 Calculation of the Standard Deviation with IBM SPSS (formerly SPSS) 68
Point of Controversy *Calculating the Sample Variance: Should We Divide by* n *or* (n − 1)? 71
5.7 Properties of the Range and Semi-Interquartile Range 72

5.8 Properties of the Standard Deviation 72
5.9 How Big Is a Standard Deviation? 73
5.10 Score Transformations and Measures of Variability 74
5.11 Standard Scores (z Scores) 74
5.12 A Comparison of z Scores and Percentile Ranks 77
5.13 Summary 77

CHAPTER 6
Standard Scores and the Normal Curve 82

6.1 Historical Aspects of the Normal Curve 83
6.2 The Nature of the Normal Curve 85
6.3 Standard Scores and the Normal Curve 86
6.4 The Standard Normal Curve: Finding Areas When the Score Is Known 87
6.5 The Standard Normal Curve: Finding Scores When the Area Is Known 90
6.6 The Normal Curve as a Model for Real Variables 92
6.7 The Normal Curve as a Model for Sampling Distributions 92
Point of Controversy *How Normal Is the Normal Curve?* 93
6.8 Summary 94

CHAPTER 7
Correlation 97

7.1 Some History 98
7.2 Graphing Bivariate Distributions: The Scatter Diagram 99
7.3 Correlation: A Matter of Direction 102
7.4 Correlation: A Matter of Degree 103
7.5 Understanding the Meaning of Degree of Correlation 105
7.6 Formulas for Pearson's Coefficient of Correlation 106
7.7 Calculating r from Raw Scores 107
7.8 Calculating r with IBM SPSS 109
7.9 Spearman's Rank-Order Correlation Coefficient 112
7.10 Correlation Does Not Prove Causation 114
7.11 The Effects of Score Transformations 115
7.12 Cautions Concerning Correlation Coefficients 116
7.13 Summary 119

CHAPTER 8
Prediction 124

8.1 The Problem of Prediction 125
8.2 The Criterion of Best Fit 127
Point of Controversy *Least-Squares Regression versus the Resistant Line* 128
8.3 The Regression Equation: Standard-Score Form 128
8.4 The Regression Equation: Raw-Score Form 130

8.5 Error of Prediction: The Standard Error of Estimate 131
8.6 An Alternative (and Preferred) Formula for S_{YX} 134
8.7 Calculating the "Raw-Score" Regression Equation and Standard Error of Estimate with IBM SPSS 134
8.8 Error in Estimating Y from X 136
8.9 Cautions Concerning Estimation of Predictive Error 138
8.10 Prediction Does Not Prove Causation 139
8.11 Summary 140

CHAPTER 9

Interpretive Aspects of Correlation and Regression 143

9.1 Factors Influencing r: Degree of Variability in Each Variable 144
9.2 Interpretation of r: The Regression Equation I 144
9.3 Interpretation of r: The Regression Equation II 146
9.4 Interpretation of r: Proportion of Variation in Y Not Associated with Variation in X 148
9.5 Interpretation of r: Proportion of Variation in Y Associated with Variation in X 149
9.6 Interpretation of r: Proportion of Correct Placements 152
9.7 Summary 152

CHAPTER 10

Probability 155

10.1 Defining Probability 156
10.2 A Mathematical Model of Probability 157
10.3 Two Theorems in Probability 158
10.4 An Example of a Probability Distribution: The Binomial 159
10.5 Applying the Binomial 162
10.6 Probability and Odds 163
10.7 Are Amazing Coincidences Really That Amazing? 163
10.8 Summary 164

CHAPTER 11

Random Sampling and Sampling Distributions 168

11.1 Random Sampling 170
11.2 Using a Table of Random Numbers 171
11.3 The Random Sampling Distribution of the Mean: An Introduction 172
11.4 Characteristics of the Random Sampling Distribution of the Mean 175
11.5 Using the Sampling Distribution of $\overline{X}$ to Determine the Probability for Different Ranges of Values of $\overline{X}$ 177
11.6 Random Sampling Without Replacement 180
11.7 Summary 181

CHAPTER 12

Introduction to Statistical Inference: Testing Hypotheses about Single Means (z and t) 183

12.1 Testing a Hypothesis about a Single Mean 184
12.2 The Null and Alternative Hypotheses 185
12.3 When Do We Retain and When Do We Reject the Null Hypothesis? 186
12.4 Review of the Procedure for Hypothesis Testing 186
12.5 Dr. Brown's Problem: Conclusion 187
12.6 The Statistical Decision 188
12.7 Choice of H_A: One-Tailed and Two-Tailed Tests 190
12.8 Review of Assumptions in Testing Hypotheses about a Single Mean 191
Point of Controversy *The Single-Subject Research Design* 192
12.9 Estimating the Standard Error of the Mean When σ Is Unknown 193
12.10 The t Distribution 195
12.11 Characteristics of Student's Distribution of t 197
12.12 Degrees of Freedom and Student's Distribution of t 198
12.13 An Example: Has the Violent Content of Television Programs Increased? 199
12.14 Calculating t from Raw Scores 201
12.15 Calculating t with IBM SPSS 204
12.16 Levels of Significance versus p-Values 206
12.17 Summary 207

CHAPTER 13

Interpreting the Results of Hypothesis Testing: Effect Size, Type I and Type II Errors, and Power 213

13.1 A Statistically Significant Difference versus a Practically Important Difference 214
Point of Controversy *The Failure to Publish "Nonsignificant" Results* 215
13.2 Effect Size 216
13.3 Errors in Hypothesis Testing 218
13.4 The Power of a Test 220
13.5 Factors Affecting Power: Difference between the True Population Mean and the Hypothesized Mean (Size of Effect) 221
13.6 Factors Affecting Power: Sample Size 222
13.7 Factors Affecting Power: Variability of the Measure 222
13.8 Factors Affecting Power: Level of Significance (α) 222
13.9 Factors Affecting Power: One-Tailed versus Two-Tailed Tests 223
13.10 Calculating the Power of a Test 224
Point of Controversy *Meta-Analysis* 225
13.11 Estimating Power and Sample Size for Tests of Hypotheses about Means 227
13.12 Problems in Selecting a Random Sample and in Drawing Conclusions 229
13.13 Summary 229

CHAPTER 14

Testing Hypotheses about the Difference between Two Independent Groups 232

14.1 The Null and Alternative Hypotheses 233
14.2 The Random Sampling Distribution of the Difference between Two Sample Means 234
14.3 Properties of the Sampling Distribution of the Difference between Means 236
14.4 Determining a Formula for *t* 237
14.5 Testing the Hypothesis of No Difference between Two Independent Means: The Dyslexic Children Experiment 239
14.6 Use of a One-Tailed Test 242
14.7 Calculation of *t* with IBM SPSS 242
14.8 Sample Size in Inference about Two Means 246
14.9 Effect Size 246
14.10 Estimating Power and Sample Size for Tests of Hypotheses about the Difference between Two Independent Means 249
14.11 Assumptions Associated with Inference about the Difference between Two Independent Means 250
14.12 The Random-Sampling Model versus the Random-Assignment Model 251
14.13 Random Sampling and Random Assignment as Experimental Controls 252
14.14 Summary 253

CHAPTER 15

Testing for a Difference between Two Dependent (Correlated) Groups 258

15.1 Determining a Formula for *t* 259
15.2 Degrees of Freedom for Tests of No Difference between Dependent Means 260
15.3 An Alternative Approach to the Problem of Two Dependent Means 261
15.4 Testing a Hypothesis about Two Dependent Means: Does Text Messaging Impair Driving? 262
15.5 Calculating *t* with IBM SPSS 265
15.6 Effect Size 267
15.7 Power 268
15.8 Assumptions When Testing a Hypothesis about the Difference between Two Dependent Means 268
15.9 Problems with Using the Dependent-Samples Design 268
15.10 Summary 270

CHAPTER 16

Inference about Correlation Coefficients 274

16.1 The Random Sampling Distribution of *r* 275
16.2 Testing the Hypothesis that $\rho = 0$ 276

16.3 Fisher's z' Transformation 277
16.4 Strength of Relationship 279
16.5 A Note about Assumptions 279
16.6 Inference When Using Spearman's r_S 279
16.7 Summary 279

CHAPTER 17
An Alternative to Hypothesis Testing: Confidence Intervals 281
17.1 Examples of Estimation 282
17.2 Confidence Intervals for μ_X 283
17.3 The Relation between Confidence Intervals and Hypothesis Testing 286
17.4 The Advantages of Confidence Intervals 286
17.5 Random Sampling and Generalizing Results 287
17.6 Evaluating a Confidence Interval 288
Point of Controversy *Objectivity and Subjectivity in Inferential Statistics: Bayesian Statistics* 289
17.7 Confidence Intervals for $\mu_X - \mu_Y$ 290
17.8 Sample Size Required for Confidence Intervals of μ_X and $\mu_X - \mu_Y$ 294
17.9 Confidence Intervals for ρ 295
17.10 Where Are We in Statistical Reform? 296
17.11 Summary 297

CHAPTER 18
Testing for Differences among Three or More Groups: One-Way Analysis of Variance (and Some Alternatives) 299
18.1 The Null Hypothesis 300
18.2 The Basis of One-Way Analysis of Variance: Variation within and between Groups 301
18.3 Partition of the Sums of Squares 302
18.4 Degrees of Freedom 305
18.5 Variance Estimates and the F Ratio 305
18.6 The Summary Table 307
18.7 Example: Does Playing Violent Video Games Desensitize People to Real-Life Aggression? 307
18.8 Comparison of t and F 310
18.9 Raw-Score Formulas for Analysis of Variance 311
18.10 Calculation of ANOVA for Independent Measures with IBM SPSS 313
18.11 Assumptions Associated with ANOVA 315
18.12 Effect Size 316
18.13 ANOVA and Power 317
18.14 *Post Hoc* Comparisons 318
18.15 Some Concerns about *Post Hoc* Comparisons 319

18.16 An Alternative to the *F* Test: Planned Comparisons 320
18.17 How to Construct Planned Comparisons 321
18.18 Analysis of Variance for Repeated Measures 323
18.19 Calculation of ANOVA for Repeated Measures with IBM SPSS 328
18.20 Summary 330

CHAPTER 19

Factorial Analysis of Variance: The Two-Factor Design 335

19.1 Main Effects 336
19.2 Interaction 338
19.3 The Importance of Interaction 340
19.4 Partition of the Sums of Squares for Two-Way ANOVA 341
19.5 Degrees of Freedom 346
19.6 Variance Estimates and *F* Tests 346
19.7 Studying the Outcome of Two-Factor Analysis of Variance 347
19.8 Effect Size 349
19.9 Calculation of Two-Factor ANOVA with IBM SPSS 350
19.10 Planned Comparisons 352
19.11 Assumptions of the Two-Factor Design and the Problem of Unequal Numbers of Scores 352
19.12 Mixed Two-Factor Within-Subjects Design 352
19.13 Calculation of the Mixed Two-Factor Within-Subjects Design with IBM SPSS 357
19.14 Summary 358

CHAPTER 20

Chi-Square and Inference about Frequencies 363

20.1 The Chi-Square Test for Goodness of Fit 364
20.2 Chi-Square (χ^2) as a Measure of the Difference between Observed and Expected Frequencies 365
20.3 The Logic of the Chi-Square Test 366
20.4 Interpretation of the Outcome of a Chi-Square Test 368
20.5 Different Hypothesized Proportions in the Test for Goodness of Fit 368
20.6 Effect Size for Goodness-of-Fit Problems 369
20.7 Assumptions in the Use of the Theoretical Distribution of Chi-Square 370
20.8 Chi-Square as a Test for Independence between Two Variables 370
20.9 Finding Expected Frequencies in a Contingency Table 372
20.10 Calculation of χ^2 and Determination of Significance in a Contingency Table 373
20.11 Measures of Effect Size (Strength of Association) for Tests of Independence 374
Point of Controversy *Yates' Correction for Continuity* 375
20.12 Power and the Chi-Square Test of Independence 377
20.13 Summary 377

CHAPTER 21
Some (Almost) Assumption-Free Tests 381

21.1 The Null Hypothesis in Assumption-Freer Tests 382
21.2 Randomization Tests 382
21.3 Rank-Order Tests 385
21.4 The Bootstrap Method of Statistical Inference 385
21.5 An Assumption-Freer Alternative to the *t* Test of a Difference between Two Independent Groups: The Mann-Whitney *U* Test 386
Point of Controversy *A Comparison of the t Test and Mann-Whitney U Test with Real-World Distributions* 390
21.6 An Assumption-Freer Alternative to the *t* Test of a Difference between Two Dependent Groups: The Sign Test 390
21.7 Another Assumption-Freer Alternative to the *t* Test of a Difference between Two Dependent Groups: The Wilcoxon Signed-Ranks Test 392
21.8 An Assumption-Freer Alternative to One-Way ANOVA for Independent Groups: The Kruskal–Wallis Test 395
21.9 An Assumption-Freer Alternative to ANOVA for Repeated Measures: Friedman's Rank Test for Correlated Samples 397
21.10 Summary 399

Epilogue: The Realm of Statistics 403

APPENDIX A
Review of Basic Mathematics 407

APPENDIX B
List of Symbols 416

APPENDIX C
Answers to Problems 420

APPENDIX D
Statistical Tables 435

Table A: Areas under the Normal Curve Corresponding to Given Values of z 436
Table B: The Binomial Distribution 440
Table C: Random Numbers 443
Table D: Student's t Distribution 445
Table E: The F Distribution 447
Table F: The Studentized Range Statistic 451

Table G: Values of the Correlation Coefficient Required for Different Levels of Significance When H_0: $\rho = 0$ 452
Table H: Values of Fisher's z' for Values of r 454
Table I: The χ^2 Distribution 455
Table J: Critical One-Tail Values of ΣR_X for the Mann-Whitney U Test 456
Table K: Critical Values for the Smaller of R_+ or R_- for the Wilcoxon Signed-Ranks Test 458

REFERENCES 459

INDEX 467

CHAPTER 1

Introduction

When you have finished studying this chapter, you should be able to:

- Define statistics as a specialization within the field of mathematics;
- Explain the difference between descriptive and inferential statistics;
- Understand that the term *statistic* has a second meaning—a descriptive index of a sample—and know the difference between a statistic and a parameter;
- Explain the role of applied statistics;
- Understand the difference between a research question and a statistical question, and the difference between a statistical conclusion and a research conclusion;
- Understand that statistical techniques are tools that are only as good as the individual who uses them, but that sometimes even good researchers make incorrect statistical conclusions;
- Understand the difference between variables and constants, and the difference between a discrete variable and a continuous variable;
- Recognize measurements as being on one of four scales—nominal, ordinal, interval, or ratio—and understand that not all numbers can be treated alike; and

- Appreciate that to learn statistics you may have to read and study differently than for other subjects.

Is statistics important to people's lives? Consider the following two research studies.

Example 1:
In early 1987, the U.S. Food and Drug Administration (FDA) was faced with an unprecedented situation. Thousands of people were dying of acquired immunodeficiency syndrome (AIDS), a relatively new sexually transmitted infection caused by the human immunodeficiency virus. The number of new cases of AIDS was increasing so rapidly that the Public Health Service was predicting that more than 50,000 Americans would soon be dying of the infection every year. Not only was there no known cure, but there was not even a drug available to slow the progression of the infection. However, early clinical trials of an experimental antiviral drug known then as azidothymidine (AZT) were promising. Only 1 of 145 AIDS patients on AZT had died, compared to 19 of 137 patients in a control group given a placebo (a substance that is medically without effect).

The process of approval of a new drug is typically very slow (8.8 years, on average), and several more years of clinical trials were scheduled. There were medical questions remaining to be answered. What was the optimal dose? For how long would the drug continue to thwart the virus? Would it prove to have toxic side effects? There were ethical questions as well. Was it morally right to withhold from other patients a drug that seemingly prolonged lives? There was also an important statistical question, one that had to be answered before the medical and ethical questions could be addressed. Was the fewer number of deaths among AIDS patients using AZT the result of the drug, or was it due just to chance? Even if the FDA had given both groups of AIDS patients the placebo or both AZT, it would be unlikely that the same number of deaths would have occurred in both groups. One would expect a difference due to chance. Were the observed differences in deaths great enough to warrant early approval of the drug? Statistical tests showed that the difference between the two groups was so great that *the probability of it having occurred by chance was less than 1 in 1,000* (Fischl et al., 1987). Armed with these statistics, the FDA gave final approval for the use of AZT in March 1987 after only 21 months of testing. The FDA's decision to give early approval for AZT was an ethical one designed to prolong lives, but it was made only after statistical analysis.

Example 2:
For women, the ovaries are a source of two important hormones: estrogen and progesterone. As women approach age 50, their supply of eggs within the ovaries nears zero, they have their last menstrual period (called menstruation), and the ovaries atrophy. The loss of estrogen and progesterone was believed to cause many serious problems (including hot flashes and weaker bone density), and thus for many years the standard medical treatment for postmenopausal women was hormone replacement therapy.

In 1997, a large study was begun to compare the long-term effects of a popular hormone replacement drug (Prempro) versus a placebo. It was expected that Prempro would have very beneficial effects in postmenopausal women. The study was originally designed to last 8 years, but it was ended after 5 years. The reason was that the women given Prempro were experiencing a greater number of heart attacks, strokes, blood clots, and breast cancers (Manson et al., 2003). Statistical analysis had shown that the greater number of bad outcomes experienced in the group given Prempro compared to the group given the placebo *had a low probability of occurring by chance* (i.e., it was more likely due to the hormone replacement therapy). The results of this study have been a major influence on how postmenopausal women receive medical treatment today.

Statistics is used not only in making important decisions, as in the cases of AZT and hormone replacement therapy, but also in our everyday lives. When your professor organizes test scores into a grade distribution and calculates the class average, he or she is using statistics. If you have ever kept track of the number of gold, silver, and bronze medals won by athletes in Olympic games competition, you have used statistics. Without statistics, data collected in our everyday observations or in carefully controlled experiments would have very little meaning.

What do we mean when we use the word *statistics*? In ordinary language, most people use the term *statistics* to refer to any set of facts that involve numbers (e.g., "birth statistics," "crime statistics," "unemployment statistics"). Here, *statistics* is a plural word (e.g., "The crime statistics are going to be reported on Wednesday"). However, in this section we refer to *statistics* as a science, a specialization within the field of mathematics. **Statistics** is the science of classifying, organizing, and analyzing data. In this sense, the word is singular (e.g., "Statistics is a science"). We introduce yet another meaning of the term *statistics* in Section 1.2.

statistics
the science of classifying, organizing, and analyzing data (for another meaning, see Section 1.2)

The goal of this book is to teach you the techniques of statistics used by researchers to analyze data. Some of these techniques will be very simple, and others will be more complex. In learning statistics, it is useful to divide the subject into two parts: descriptive statistics and inferential statistics.

1.1 Descriptive Statistics

In a new school, a biology instructor is thinking about what the first group of students will be like. How much biological information will they already have? She does not wish to bore them by underestimating their knowledge, but she also does not want to lose them by assuming too much. Should she begin with a review of fundamentals? How extensive? In what areas? It would also be helpful to know whether the students vary widely in the extent of their biological knowledge. If they do, the instructional method may need adjustment.

Let us suppose that the biology instructor administers a nationally developed test of biological knowledge to her students at the first class meeting. Let us also assume that she is able to learn how students in similar educational circumstances perform on this test. With this knowledge, she discovers, for example, that Bob Beaver's score in zoology is better than that of 90% of students in the same grade and that only 15% of students get lower scores than Steve Smith's.

The biology instructor finds that there are too many scores to keep in mind at once. Because the names in her gradebook are in alphabetical order, the scores are not in numerical order, so she first organizes the scores in descending order and then finds the class average for each subtest and for the complete test. Then she compares the class averages with the performance data for other similar students. She finds that as a group her students are approximately at the expected level in botany and that their performance is superior in zoology. This knowledge will help her go about teaching.

In reorganizing and comparing her students' scores, our instructor is making use of techniques that are part of the body of descriptive statistics. The purpose of **descriptive statistics** is to organize and to summarize observations so that they are easier to comprehend.

descriptive statistics
its purpose is to organize and summarize observations

1.2 Inferential Statistics

What is the attitude of the voting public toward capital punishment? A second branch of statistical practice, known as inferential statistics, provides the basis for answering questions of this kind. Pollsters would find it impossible to ask all registered voters a question like the one about capital punishment, so they ask only a small portion of voters and then use the answers to estimate the attitudes of all voters. The purpose of **inferential statistics** is to draw a conclusion (an inference) about conditions that exist in a **population** (the complete set of observations) by studying a **sample** (a subset) drawn from the population. Because of chance factors associated with drawing a sample, the outcome, like any estimate, is subject to error. However, if the sample of voters selected for study has been chosen according to statistical principles, it is possible to know what margin of error is involved.

We previously defined statistics as a science. However, the term statistic has another meaning that we will use throughout the book. A **statistic** is a descriptive index of a sample. The same index, if descriptive of a population, is called a **parameter**. Thus, the average of a sample of scores is a statistic; the average of a population of scores is a parameter. *In inferential statistics, we will be interested in estimating population parameters from sample statistics.*

Here is another example of how inferential statistics is used to evaluate the outcome of an experiment. Is it possible that a certain drug has an effect on the speed of learning? Let us suppose that an investigator decides on the kind of subjects she wishes to study. As in the first example, it is impossible to administer the drug to everyone in the population, so the investigator selects at random two samples of 25 subjects each. **Random sampling** is a procedure that ensures that all samples of a particular size have an equal chance of being selected and thus eliminates any bias when we draw a sample. The investigator then administers the drug to one of the groups and a placebo to the other group. She gives both groups a learning task and treats both alike in all ways. She finds that the average learning scores of the two groups differ by 5 points.

We would expect some difference between the groups even if both received the drug because of chance factors involved in the random selection of the groups. The question that the experimenter must ask herself is whether the observed difference is within the limits of expected variation. If certain preconditions have been met, statistical theory can provide the basis for an answer. If the experimenter finds that the obtained difference of 5 points is larger than can be accounted for by chance variation, she will infer that other factors must be at work. If examination of her experimental procedures reveals no reasonable cause for the difference other than the deliberate difference in the experimental treatment, she may make an inference about the effects of the drug in the population. This is the same type of reasoning that the FDA investigators used to conclude that AZT prolonged the lives of AIDS patients. Their conclusion was not about just those patients in the study. It was about the effects of AZT if it were given to *all* persons with AIDS.

inferential statistics
its purpose is to draw a conclusion about conditions that exist in a population from study of a sample

population
the complete set of observations about which an investigator wishes to draw conclusions

sample
a subset of a population

statistic
a descriptive index of a sample (for another meaning, see page 3.)

parameter
a descriptive index of a population

1.3 Our Concern: Applied Statistics

Statistics originated from the work of mathematicians and others who were well grounded in mathematics. However, the mathematical development of statistics is not what this book is about. Instead, we will be concerned with applied statistics. Our outlook will be that of investigators who need to know statistics to appreciate reports of findings in their professional fields or who must apply statistical treatment in their own work. We may think of ourselves as statistical technicians, therefore. Among our ranks are biologists, educators, psychologists, engineers, sociologists, medical researchers, and business executives. All these professionals and many more regularly find that statistical procedures can be helpful in their work.

Applied statistics is a tool and neither a beginning nor an end in itself. For any subject area, scientific study begins with an investigator posing a question based on the results of previous studies and theoretical considerations. This is the **research question**, a question of fact concerning the subject matter under investigation. Suppose, for example, that we are interested in whether reaction time depends on the intensity of a stimulus.

We must next decide on the specifics necessary to explore the question. We may decide to assign 50 available subjects at random to two groups of 25, one of which will receive a

random sampling
a procedure that ensures that all samples of a particular size have an equal chance of being selected

research question
a question of fact about the subject matter under investigation

strong stimulus and the other a weaker one. Other designs are possible, but the experimental design will depend, for the most part, on the research question. For example, if we believe that reaction time decreases as stimulus intensity increases, we could test reactions to more than two intensities of the stimulus and assign fewer subjects to a larger number of groups.

The variable that is systematically manipulated by the investigator is called the **independent variable** (e.g., stimulus intensity). The variable that is measured is called the **dependent variable** (e.g., reaction time). It is important that all other factors that might affect the dependent variable (**confounding variables**) be the same for all groups. For example, it would be important to see that the subjects have been tested at approximately the same time of day to avoid having one group fresh and alert while the other is weary and inattentive. Having designed our study, we are finally ready to make observations. We measure the reaction time of each subject and record it.

After recording our measurements, we calculate the average reaction time for each group, see that they differ, and raise a statistical question: Is the average reaction time under the two conditions so different that chance variation alone cannot account for it? A **statistical question** is a question about a numerical aspect of the observations. After applying a statistical procedure, we arrive at a statistical conclusion. A **statistical conclusion** is a conclusion about a numerical property of the data. For example, as a result of applying the appropriate statistical procedures in our hypothetical experiment, we might be able to conclude that the average reaction times under the two conditions are so different that it is not reasonable to believe that chance alone could account for it.

Now we return to the research question. If we had carefully arranged the conditions of the study, it may be possible to conclude that intensity of the stimulus does make a difference, at least under the conditions that held in this experiment. This is our final step, and it is a research conclusion. A **research conclusion** is a conclusion about the subject matter. In our initial example about AZT, the statistical conclusion was that the number of deaths in the two groups differed by beyond what would have been expected by chance. The research conclusion was that the difference was due to AZT. Although the research conclusion derives partly from the statistical conclusion, we see that other factors must be considered. The investigator, therefore, must weigh both the statistical conclusion and the adequacy of the research design in arriving at the research conclusion.

It is important to understand that **statistical procedures** *are only a middle step in the investigation of a research question.* If an investigator has not carefully controlled for all confounding variables, it is possible to make a correct statistical conclusion and at the same time make an incorrect research conclusion. Such was the case in a well-known study by Joseph Brady and some colleagues at Johns Hopkins University (Brady, Porter, Conrad, & Mason, 1958). They investigated the effects of stress on the development of ulcers in four pairs of monkeys. Each pair was locked into adjacent chairs during testing. One monkey in each pair, designated the "executive" monkey, could avoid a strong electric shock by pressing a lever that postponed the shock for 20 seconds. Whenever the executive monkey did receive a shock, the other monkey in the pair, called the "yoked" control monkey, also received a shock (which it could not avoid). The investigators assigned monkeys to the executive group based on their high rates of lever pressing in previous studies.

The executive monkeys developed ulcers, but the yoked monkeys did not, even though they received identical shocks. The difference was so great that it was unlikely that the results occurred by chance (the statistical conclusion). Because they believed that they had controlled for all other factors, Brady and his colleagues concluded that it was the continuous stress of having to decide how often to press the lever to avoid shock that caused the ulcers in the executive monkeys (the research conclusion).

Later studies were unable to replicate the results. Moreover, it was found that animals selected for high rates of responding were more susceptible to developing ulcers than were other animals (Weiss, 1971). When monkeys that did not differ in activity level were tested together, the executive monkeys developed less severe ulcers (Foltz & Millett, 1964; Natelson, 1976).

independent variable
the variable that is systematically manipulated by the investigator

dependent variable
the variable that is measured

confounding variables
factors (other than the independent variable) that might affect the dependent variable

statistical question
a question about a numerical aspect of the observations

statistical conclusion
a conclusion about a numerical property of the data

research conclusion
a conclusion about the subject matter

statistical procedures
only a middle step in the investigation of a research question

We now know that in general, control over an aversive event reduces the stress the event causes (Taylor, 1991). The original investigators had unknowingly biased the outcome of their experiment by not randomly assigning the monkeys to the two conditions and thus reached an incorrect research conclusion.

1.4 Variables and Constants

variable
a characteristic that may take on different values

When we conduct a study, we must often consider several variables and constants. A **variable** is a characteristic that may take on different values. Typical examples are intelligence test scores, number of errors on a spelling test, height, marital status, and gender. The concept of a variable does not imply that each observation must differ from all the others. All that is necessary is the possibility of difference. Suppose, for example, that a school nurse is interested in the height of seventh-grade boys in Lincoln Junior High School. If she selects a sample of three for study and finds that they are all the same height, it is still proper to refer to height as a variable because the possibility of getting students of different height existed.

constant
a characteristic that can have only one value (other values are not possible)

On the other hand, her decision to study height only among the seventh-grade students means that grade level is a constant rather than a variable in this inquiry. When, in terms of a particular study, it is not possible for a characteristic to have other than a single value, that characteristic is a **constant**. Constants limit the applicability of the results of a study. In the school nurse's situation, for example, the sex of the students (male), the school (Lincoln), and the grade level (seventh) are all constants. Even if the nurse had taken a sample of larger size, the conclusions that she would draw could apply for sure only to seventh-grade boys in that school.

discrete variable
a variable that can take on only certain values

Variables may be either discrete or continuous. A **discrete variable** can take on only certain values. If we count the people at a meeting, we might learn that 73 persons attended, or 74, but no value between these two figures is possible. The values are *exact numbers.* Other examples include the number of multiple-choice questions answered on an exam or the number of lever presses performed by a rat in a learning experiment.

continuous variable
a variable that can take on any value (within the limits that its values may range)

A **continuous variable** can take on any value (within whatever limits its values may range). In measuring length, for example, it is possible for an object to be 3 ft. 1 in. or 3 ft. 2 in. long, or any conceivable length in between. A continuous variable has no gaps in its scale. Other examples include weight, age, and temperature.

Even though a variable is continuous in theory, the process of measurement always reduces it to a discrete one. For example, if we measure weight to the nearest tenth of a pound, a weight of 18.3 pounds means that the object is closer to 18.3 pounds than it is to 18.2 pounds or 18.4 pounds. The value of 18.3 pounds is an *approximate number.* The actual weight is somewhere between 18.25 pounds and 18.35 pounds. In this example, our recorded measurements form a discrete scale in steps of one-tenth of a pound. When measuring a continuous variable, our level of accuracy is limited by our recording equipment and it is up to the investigator to determine the degree of precision appropriate to the problem at hand. Your height may be 67 and 7/16 in., but who cares? For most of us, measurement to the nearest half-inch is fine.

Typical statistical computations involve both exact and approximate numbers—for example, dividing the sum of the scores (usually the sum of approximate numbers and thus itself approximate) by the number of cases (an exact number) to produce an average. *In computations involving both exact and approximate numbers, the accuracy of the answer is limited by the accuracy of the approximate numbers involved*. For example, if we measure weight to the nearest tenth of a pound and find that the total weight of three objects is 67.0 pounds, it would be misleading to report the average weight (67.0/3) as either 22.333333 pounds (the output of the common eight-digit display calculator) or 22 pounds. Because the approximate number is reported to the nearest tenth of a pound, the reported average weight should be 22.3 pounds.

When working with a series of statistical computations involving approximate numbers, it is a good idea to keep a little more accuracy than we think is the minimum to which we are entitled (perhaps one extra decimal place) because in a sequence of computations it is possible to compound inaccuracy. Once the computation is completed, we should round back to the least accurate of the numbers used. If the last digit is greater than 5, round up; if it is less than 5, round down. *When rounding a number that ends exactly with the numeral 5, you should always round to the nearest even number.* This avoids introducing a bias in your calculations. (Sometimes you will round up and other times down rather than always in the same direction.) For example, if rounding to the nearest tenth, 1.35 becomes 1.4, as does 1.45.

1.5 Scales of Measurement

measurement the process of assigning numerals to observations

Measurement is the process of assigning numerals to observations. This is not done arbitrarily, but in a way so that the numbers are meaningful. How do we do this? Variables do not all have the same numerical properties. In measuring weight, we are accustomed to the idea that 40 pounds is twice as much as 20 pounds, and that the difference between 10 and 20 pounds is the same as that between 60 and 70 pounds. Similarly, if a baseball team has won 40 games and another only 20, we can say that the first has won twice as many games as the second team. However, if we put baseball teams in order of their standing in the league, we should not think that the number one team is twice as good as the number two team, nor would we expect that the difference in abilities between the first and second team is necessarily the same as that between the second and third team. It is apparent that numbers have a different significance in these two situations. To help distinguish different kinds of situations, the psychologist S. S. Stevens (1946) identified four different scales of measurement. You will find that numbers are treated differently depending on their scale of measurement.

Nominal Scale

nominal scale mutually exclusive and exhaustive categories differing in some qualitative aspect

Some variables are qualitative in their nature rather than quantitative. For example, eye color, types of cheese, and party of political affiliation are examples of *qualitative variables*. The several categories of such a variable form a **nominal scale** (the Latin root of *nominal* means "name"). The only requirements of a nominal scale are that the categories be *mutually exclusive* (the observations cannot fall into more than one category) and *exhaustive* (there must be enough categories for all the observations). Brown and blue eyes are mutually exclusive categories for eye color; but Ford, General Motors, Toyota, Chevrolet, and Honda are not mutually exclusive categories for types of cars (Chevrolets are made by General Motors). With a nominal scale, there is no question about one category having more or less of any particular quality; all categories are simply different. We may use numbers to identify the categories of a given scale. For example, we might identify three cheeses as Cheese No. 1, Cheese No. 2, and Cheese No. 3. Numbers used in this way are simply a substitute for names and serve only for purposes of identification. Numbers on football jerseys similarly indicate only the position of the player (e.g., offensive line, offensive backfield). If one football player wears the number 10 on his back while another wears the number 20, there is no implication that the second player is "more than" the other in some dimension, let alone that he has twice as much of something (Stine, 1989).

Ordinal Scale

At the next level of complexity is the **ordinal scale** (the Latin root means "order"). In this type of measurement, *the categories must still be mutually exclusive and exhaustive, but they also indicate the order of magnitude of some variable.* With a nominal scale, the outcome of classification is a set of unordered categories. With the ordinal scale, it is a set of ranks. A classic example

ordinal scale has the properties of a nominal scale, but in addition the observations may be ranked in order of magnitude (with nothing implied about the difference between adjacent steps on the scale)

is military rank. Sergeant is greater than corporal, which is in turn greater than private. Another example is academic standing—freshman, sophomore, junior, and senior. In yet another example, a supervisor may estimate the competence of seven workers by arranging them in order of merit. The only relation expressed in a series of numbers used in this way is that of "greater than." We may use numbers for the ranks but it is not necessary. Among persons ranked 1, 2, and 3, the first person has a greater degree of merit than the person ranked second, and the second person has greater merit than the third. However, nothing is implied about the magnitude of difference in merit between adjacent steps on the scale. Furthermore, nothing is implied about the absolute level of merit; all seven workers could be excellent, or they could be quite ordinary.

Interval Scale

interval scale has all the properties of an ordinal scale, and a given distance between measures has the same meaning anywhere on the scale

The next major level of complexity is the **interval scale**. This scale has all the properties of the ordinal scale, but with the further refinement that a *given interval (distance) between scores has the same meaning anywhere on the scale*. Thus, it could be better called an *equal-interval scale*. Examples of this type of scale are degrees of temperature on the Fahrenheit or Celsius scales. A 10° rise in a reading on the Celsius scale represents the same change in heat when going from 0° to 10° as when going from 20° to 30°. The limitation of an interval scale is that it is not possible to speak meaningfully about a ratio between two measurements. We illustrate this point in Figure 1.1. The top of this illustration shows three temperatures in degrees Celsius: 0°, 50°, and 100°. It is tempting to think of 100°C as twice as hot as 50°. However, the value of zero on this scale is simply an arbitrary reference point (the freezing point of water) and does not imply an absence of heat. Therefore, it is not meaningful to assert that a temperature of 100° Celsius is twice as hot as one of 50° or that a rise from 90° to 99° Celsius is a 10% increase.

Ratio Scale

ratio scale has all the properties of an interval scale plus an absolute zero point

A **ratio scale** possesses all the properties of an interval scale and in addition *has an absolute zero point*. The bottom of Figure 1.1 shows the same three temperatures in degrees Kelvin. This scale uses the same unit for its intervals; for example, 50° of change in temperature is the same on both scales. However, the Kelvin scale has an absolute zero, the point at which a substance would have no molecular motion and, therefore, no heat. Thus, 100° is twice as hot as 50° on the Kelvin scale (100° is only 1.15 times as hot as 50° on the Celsius scale). Other examples of ratio scale measurements are length, weight, and measures of elapsed time. Not only is the difference between 40 in. and 41 in. the same as the difference between 80 in. and 81 in., but it is also true that 80 in. is twice as long as 40 in.

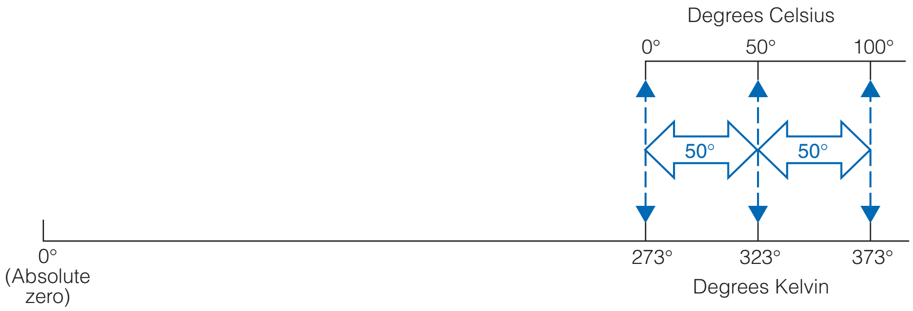


FIGURE 1.1 Three temperatures represented on the Celsius and Kelvin scales.

TABLE 1.1 *Scales of Measurement and Their Characteristics*

SCALE	PROPERTIES	EXAMPLES
Nominal	Mutually exclusive and exhaustive categories differing in some qualitative aspect.	Sex, ethnic group, religion, eye color, academic major
Ordinal	Scale has the property of a nominal scale (mutually exclusive categories) and in addition has observations ranked in order of magnitude. Ranks, which may be numerical, express a "greater than" relationship, but with no implication about how much greater.	Military rank, academic standing, workers sorted according to merit
Interval	Scale has all the properties of an ordinal scale, and in addition, numerical values indicate order of merit and meaningfully reflect relative distances between points along the scale. A given interval between measures has the same meaning at any point in the scale.	Temperature in degrees Celsius or Fahrenheit
Ratio	Scale has all the properties of an interval scale, and in addition has an absolute zero point. Ratio between measures becomes meaningful.	Length, weight, elapsed time, temperature in degrees Kelvin

Table 1.1 summarizes the characteristics of the four scales of measurement we have just discussed.

1.6 Scales of Measurement and Problems of Statistical Treatment

An understanding of the scales of measurement provides a framework for appreciating some problems in the interpretation of data. The first thing we need to realize is that in the behavioral sciences, there are many measuring instruments that lack equal intervals and an absolute zero point. Consider, for example, a spelling test. Easier items might be words such as *garden*, *baseball*, and *rowboat*. But we might also find *perceive, complacency*, and perhaps even some stumpers like *gauge*, *accede*, and *abscissa*. A score of zero on this test means that the person could not spell the simplest word on the list, but what if simpler words had been on the test, such as *cat*, *run*, and *bat*? Our spelling test, then, does not have an absolute zero point because zero on the spelling test does not indicate a total absence of spelling ability. The same is true of midterm tests, IQ tests, the SAT, and almost all other tests of mental performance.

What about equal intervals? To have equal intervals on our spelling test, we should be able to state quantitatively just how much more spelling ability is needed to spell *garden* than to spell *cat* and how much more is needed to spell *gauge* than to spell *rowboat*. So far, we have found no objective way to determine when equal numerical intervals on a mental test represent equal increments in performance. Realization of this situation once caused great concern in statistical circles. Some people argued that calculating certain statistical variables (such as averages) on tests of mental abilities could be seriously misleading. Fortunately, the weight of the evidence suggests that in most situations, making statistical conclusions is not seriously hampered by uncertainty about the scale of measurement. We will consider this matter further in the Point of Controversy in Chapter 4.

There are, however, several areas where we need to be aware of scale problems to avoid taking tempting but erroneous positions. For example, we should not say that a person with an IQ of 150 is twice as bright as one with an IQ of 75, or that the difference between 15 and 25 points on a spelling test necessarily represents the same increment in spelling ability as the difference between a score of 30 and 40 points on the same test. In psychological measurement, this problem may be particularly critical when a test does not have enough "top" or "bottom" to differentiate adequately among the group measured. For example, imagine a test of ability that has a maximum possible score of 50 points and that is too easy for the group measured. For two persons who score 50 points, the score for one may indicate the maximum level of achievement, but the second person may be capable of a much higher level of performance. The measuring instrument is simply incapable of showing this difference because it does not include items of greater difficulty.

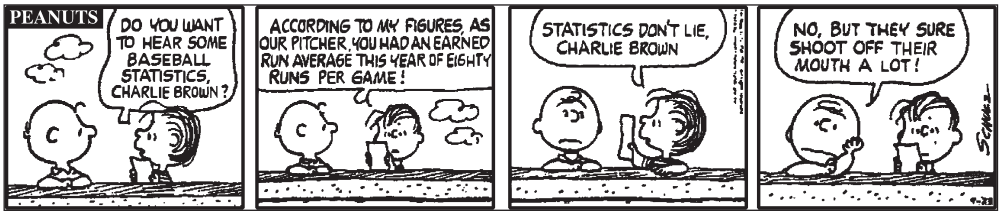
1.7 Do Statistics Lie?

> **There are three kinds of lies: lies, damned lies, and statistics.**
> Benjamin Disraeli, 19th-century British statesman

Do statistics really lie? In the experiment by Brady and his colleagues, the investigators reached an erroneous research conclusion, but there really was an observed difference between the executive and yoked monkeys in their susceptibility to developing ulcers. Is it possible to make an incorrect statistical conclusion which, of course, would lead to an invalid research conclusion? The answer is "yes." You may have noted in our previous examples that the statistical conclusions always made some statement about whether or not the observed result could have occurred by chance. We will discuss this in greater detail later, but accept for now that a statistical conclusion is always about a probability. When the average scores for two groups are found to be "statistically different," it means only that the numerical results had a low probability of occurring by chance alone. However, even events with a low probability (getting struck by lightning, for example) do occur sometimes, and when that happens with empirically obtained data, it results in an incorrect statistical conclusion.

Let us give an example. In the 1960s, a few investigators published results of studies in which they had trained planaria (very small flatworms) on a simple task, then sacrificed them and fed their brains to naive (untrained) planaria. The "cannibal" planaria's subsequent average performance on the same task was much better than one would reasonably expect to have occurred by chance (the statistical conclusion). In fact, the results had only a 5% (1 in 20) chance or less of having occurred only by chance. The investigators concluded that memory was stored in the brain in discrete chemical units called engrams that the cannibal planaria had consumed, resulting in their "knowing" how to perform the task (the research conclusion).

Scientific journals generally publish only results that would have occurred by chance 5% of the time or less. As it turns out, many other scientists had attempted the same experiment and had not observed any improvement in performance by the cannibal planaria that was so unlikely that it could not have occurred by chance. Therefore, they had not submitted their results for publication. When they talked to each other and learned of each other's failures, 23 of them got together and published a rebuttal (Byrne et al., 1966). Even though there is apparently no chemical engram, with so many investigators attempting the experiment, it was


PEANUTS
DO YOU WANT TO HEAR SOME BASEBALL STATISTICS, CHARLIE BROWN?
ACCORDING TO MY FIGURES, AS OUR PITCHER, YOU HAD AN EARNED RUN AVERAGE THIS YEAR OF EIGHTY RUNS PER GAME!
STATISTICS DON'T LIE, CHARLIE BROWN
NO, BUT THEY SURE SHOOT OFF THEIR MOUTH A LOT!

PEANUTS: © United Feature Syndicate, Inc.

to be expected that 1 in 20 would observe an average numerical difference between cannibal planaria and other planaria that would have occurred by chance only 5% of the time. It is a good thing that the original investigators' research conclusion was wrong. Otherwise, there would be a far easier way for students to learn statistics from their professor than having to sit through a course.

Although an incorrect conclusion was made in both the executive-monkeys and cannibal-planaria experiments, there was no intent on the part of the investigators to deceive anyone. The incorrect conclusions were honest mistakes. Unfortunately, there have also been cases of scientific fraud. Some claim that one of the most notorious was that of Cyril Burt (Hearnshaw, 1979; MacKintosh, 1995; Samelson, 1996), who for years was recognized as one of the leading investigators of intelligence. Burt was a strong proponent of the view that intelligence is biologically determined. His numerous published results of similar IQ scores in identical twins raised apart provided the strongest evidence for this viewpoint. After his death, it was discovered that not only had Burt fabricated some of his data, but that he may have put the names of fictitious coauthors on his papers as well. Burt is certainly not the only one guilty of scientific fraud. In 2005, the prestigious journal *Science* had to retract two papers by researcher Woo-Suk Hwang when it was learned that he had fabricated his results with stem cell cloning. The statistical procedures used to analyze data cannot tell whether the numbers used are the result of empirical observations or someone's imagination. We might remember the old saying: "Figures never lie, but liars often figure."

Sometimes an individual can mislead others not by fabricating data but by presenting only part of them. For example, in a given state college, it may be that 80% of the courses are taught by instructors holding doctorate degrees, but that only 60% of all the instructors at the institution hold the degree. If a journalist's theme is the "rotten state of higher education," we can guess which figure is more likely to be mentioned. On the other hand, if an economy-minded representative of the taxpayers' association wishes to argue that higher salaries are unnecessary to attract a fully qualified faculty, a different choice will be made. Unfortunately, if we are presented with only one of these approaches, the possibility of others may never occur to us.

The misuse of statistics is an important topic, but limitations of space prevent a more detailed treatment here. For those who are interested, we refer you to Huff (1993) and Best (2001).

Point of Controversy

Are Statistical Procedures Necessary?

In a 1988 article, psychologist Robert Bolles makes a spirited attack against the use of statistics. He advises that "one should avoid statistics whenever possible . . . and get on with the business of science." Specifically, he suggests that scientists should measure things rather than test hypotheses. They should look for the obvious. (When you see it, why do you need statistics?) And they should look at the data rather than the statistics. (Your data "will tell you if you have found something, even while statistical tests are fibbing, lying, and deceiving you.") Another researcher has argued that a good graph is worth a thousand statistical tests (Loftus, 1993) and more recently others have proposed that graphs can provide an alternative to statistical significance tests (Smith et al., 2002). Rather than consult a statistician, Bolles suggests that we "talk to Buddha or somebody like that," for if "one of the gods wants you to discover, you certainly do not have to do a statistical test."

Bolles is not the only researcher to reject statistics. B. F. Skinner, Murray Sidman, and many others who study operant conditioning (experimental analysis of behavior) abhor averaging data and the use of inferential statistics (e.g., Hopkins, Cole, & Mason, 1998; Perone, 1999). In fact, many of the studies included in Hock's (2006) *Forty Studies That Changed Psychology* are case studies (e.g., Freud, Watson, Harlow) without any statistics.

Bolles gives several examples from his own research in which there were near-perfect relationships between variables or where some experimental manipulations produced 100% change. In cases like these, statistics are, of course, superfluous. However, this does not mean that we should abandon statistics, for to do so would relegate us to areas of research where only large effects occur. At what point do you draw the line between that which is obvious and that which is not so obvious? Statistical procedures give us a way of answering this without resorting to complete subjectivity.

The criticisms by Bolles and others are not totally without merit. There are those who use statistics to lend apparent importance to trivial results. Statistics cannot tell whether a question is scientifically important or of practical use. As Cohen (1988) emphasizes, statistical techniques "can never produce high quality science in the absence of thorough consideration of basic questions."

> **Statistics are no substitute for judgment.** Henry Clay, American statesman, 1777–1852

Another problem is that some researchers try to impress others by using very complex statistical analyses when much simpler analyses would suffice. As stated by Christopher Peterson (2009), "Psychology would be improved if researchers stopped using complicated designs, procedures, and statistical analyses for the sole reason that they are able to do so." This point was also made by the American Psychological Association's Task Force on Statistical Inference (1999):

> **Although complex designs and state-of-the-art methods are sometimes necessary to address research questions effectively, simpler classical approaches often can provide elegant and sufficient answers to important questions. Do not choose an analytic method to impress your readers or to deflect criticism** (Wilkinson et al., 1999).

As you read this book, keep in mind that *statistical techniques are tools* that are only as good as the individual who uses them. With good questions at the beginning of a study and good common sense throughout, statistical procedures are very helpful tools in arriving at answers to those questions.

1.8 Some Tips on Studying Statistics

> **A good teacher is a tough guy who cares very deeply about something that is hard to understand.** Norman MacLean, University of Chicago

Is statistics a hard subject? It is and it isn't. In general, learning how to do it requires attention, care, and arithmetic accuracy, but it is not particularly difficult. Learning the "why" of things may be harder. Look again at the title of your textbook. The emphasis in this book is *reasoning* (the logic of statistics), not mathematical formulas and proofs.

What about the expected reading rate for a book about statistics? A 4-page assignment in mathematics may require the same time as 40 pages in history. Certainly, one should not expect to read a statistics text like a novel, or even like the usual history text. Don't feel stupid if you cannot race through a chapter and instead find that some absorption time is required. Most chapters will require several readings.

Although the emphasis of this book is on reasoning, you will, of course, be required to work mathematical problems. Like many students, you may have anxieties about mathematics. However, you will need to know only ordinary arithmetic and simple algebra. For students who feel at home in elementary algebra, we include algebraic notes at the end of several chapters. They provide a supplementary "handle" to better understand statistical techniques. On the other hand, we provide a review of elementary mathematics in Appendix A for those who feel that their mathematics lies in the too distant past to ensure a sense of security.

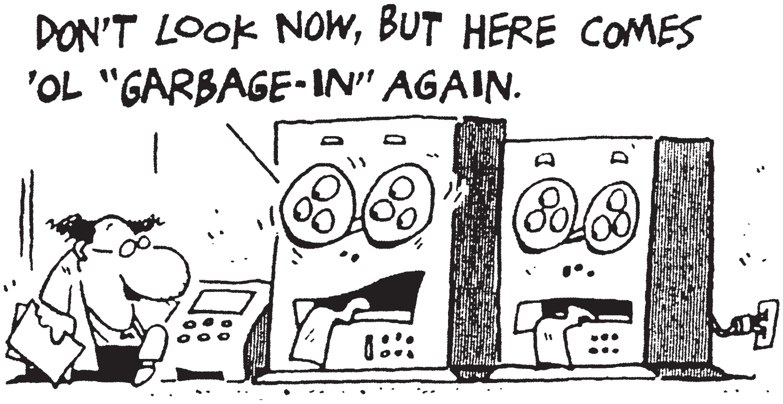
We include questions and problems at the end of each chapter and suggest that you work enough of them to feel comfortable. You will benefit by trying to answer all questions and problems, even though your teacher may assign only a few. When working problems, try constantly to see the big picture. Look at the result of each calculation. Does it make sense? You should be suspicious if you find the average to be 53 but most of the numbers are in the 60s and 70s. Remember Minium's First Law of Statistics: "*The eyeball is the statistician's most powerful instrument.*"

We need to say a word about the ladder-like nature of a course in elementary statistics. What you learn in earlier stages becomes the foundation for what follows. Consequently, it is very important that you keep up. If you have difficulty at some point, seek assistance from your instructor. Don't delay. Those who think matters may clear up if they wait may be right, but the risk is greater than in courses where the material is less interdependent. It can be like trying to climb a ladder with some rungs missing. Cramming usually does not work in statistics. Success in studying statistics depends on regular work.

1.9 Statistics and Computers

Today, computers and software for statistics have greatly eased the computations necessary to solve many statistical questions. In fact, your course may require that you learn how to use computerized statistical packages. There is a practical advantage to this real-life application. However, the emphasis of this book is on understanding statistics. Without a basic understanding of statistics, computer programs can be misused. Suppose, for example, an individual enters his or her data incorrectly. Garbage in always results in garbage out. Without a basic understanding of statistics, he or she will just accept the printout as correct. Another person knowledgeable in statistics will have some idea of what the result should look like and will check the way the data were entered. We know of a few cases, for example, in which investigators have ignored the obvious in their data when statistical results, done by computer, indicated something else. "After all," they insisted, "computers don't lie." With some difficulty, the investigators were persuaded to recheck their data, only to discover that they had entered them in the computer incorrectly. The American Psychological Association's Task Force on Statistical Inference cautions:

> "As soon as you have collected your data, before you compute *any* statistics, *look at your data*. . . . More important than choosing a specific statistical package is verifying your results, understanding what they mean, and knowing how they are computed. . . . Do not report statistics found on a printout without understanding how they are computed or what they mean. . . . There is no substitute for common sense" (Wilkinson & the Task Force on Statistical Inference, 1999).


DON'T LOOK NOW, BUT HERE COMES 'OL "GARBAGE-IN" AGAIN.

Frank & Ernest © NEA Reprinted by Permission.

In this book, we will introduce you to one of the most popular of the many statistical software packages: IBM SPSS. IBM SPSS was formerly known as SPSS (Statistical Package for the Social Sciences), the name of the company that markets and sells the software. SPSS was acquired by IBM in October 2009.

If you have experience with computers, you will probably find the IBM SPSS interface to be similar to other types of software that you have used. For example, similar to Microsoft Excel, IBM SPSS has point and click features, drop down menus, dialog boxes, and help buttons. Thus, it shouldn't take you long to familiarize yourself with this new software.

We will introduce IBM SPSS to you in Chapter 5 with calculation of the standard deviation. As in all the subsequent chapters, use of IBM SPSS will be demonstrated only after you have learned how to work the problems by hand (and calculators) with formulas. You will find that a problem that may take an hour to work by hand will take only a couple of minutes with IBM SPSS. However, remember again that there is no substitute for *understanding* conceptually and mathematically a statistical technique.

1.10 Summary

Statistics is the science of classifying, organizing, and analyzing data. There are two main branches: descriptive statistics and inferential statistics. The techniques of descriptive statistics allow us to organize and summarize observations so they become easier to understand. Procedures in inferential statistics are used to draw inferences about conditions that exist in a population from study of a sample (a subset of the population). The word *statistic* has a second meaning—a descriptive index of a sample. A descriptive index of a population is called a parameter.

It is important to distinguish between research matters and statistical matters. Research questions and conclusions concern the subject matter under investigation, whereas statistical questions and conclusions apply to the numerical characteristics of the data. Statistical procedures are only a middle step in arriving at a research conclusion. The typical steps of an investigation are

Research question → Statistical question → Collection of data → Statistical conclusion → Research conclusion

A variable is a characteristic that may take on different values. Constants, on the other hand, remain the same in a given study. Discrete variables have exact values, like the number of subjects in an experiment. Continuous variables can take on any value, but the process of measurement always reduces them to discrete variables. Thus, measurement of a continuous variable gives us approximate values, like the length or weight of an object. When computations involve both approximate and exact numbers (as is typical in calculating an average), the approximate numbers dictate the accuracy of the outcome.

Variables may be qualitative, expressing only differences in kind (nominal scale), or they may be quantitative, expressing differences in amount (ordinal, interval, and ratio scales). Numbers can be used to identify different values on any of the scales, but the meaning of these numbers differs with the nature of the scale. Most tests in the behavioral sciences lack equal intervals and an absolute zero point. Although these shortcomings do not ordinarily prevent us from using statistical procedures, scale problems can lead to incorrect research conclusions.

Although statistical theory is rooted in mathematics and probability, it is quite possible to acquire a good working understanding of elementary statistics without having to take advanced courses in math. You will only have to do ordinary arithmetic and follow what is happening in simple equations. The emphasis in this book will be on conceptual development.

Key Terms, Concepts, and Symbols

statistics (p. 3)
descriptive statistics (p. 3)
inferential statistics (p. 4)
population (p. 4)
sample (p. 4)
statistic (p. 4)
parameter (p. 4)
random sampling (p. 4)
research question (p. 4)
independent variable (p. 5)
dependent variable (p. 5)
confounding variables (p. 5)
statistical question (p. 5)
statistical conclusion (p. 5)
research conclusion (p. 5)
statistical procedures (p. 5)
variable (p. 6)
constant (p. 6)
discrete variable (p. 6)
continuous variable (p. 6)
measurement (p. 7)
nominal scale (p. 7)
ordinal scale (p. 8)
interval scale (p. 8)
ratio scale (p. 8)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. An experimenter may "wear two hats," that of subject-matter expert and that of statistician. Is our experimenter wearing primarily the first hat, primarily the second, or both about equally when he or she: (a) Thinks up the problem? (b) Translates it into a statistical question? (c) Draws a conclusion that the average performance of experimental and control groups is really different? (d) Decides that the imposed difference in treatment was responsible for the difference in average performance? (e) Relates this finding to those of previous studies?

2. The subject of statistics is often divided into two parts: descriptive and inferential. Briefly describe each.

3. For the following examples, is the individual primarily using descriptive statistics or inferential statistics? (a) A pollster asks a group of voters how they intend to vote in the upcoming election for governor. (b) A researcher tests a new diet drug on a group of overweight individuals. (c) A teacher organizes test grades into a distribution, from best to worst. (d) A sports fan ranks a team's players according to their batting averages.

4. We are interested in the choices of majors of this year's entering freshmen at our university. We randomly survey 10% of them. (a) What is the population? (b) What is the sample? (c) What is the parameter? (d) What is the statistic?

5. Is it possible to draw a correct statistical conclusion yet reach an incorrect research conclusion? Explain your answer.

6. Is it likely to arrive at a correct research conclusion after making an incorrect statistical conclusion? Explain your answer.

7. If the effect of the act of measurement is disregarded, which of the following variables are best regarded fundamentally as forming a discrete series, and which a continuous series? (a) Temperature. (b) Time. (c) Sex. (d) Brands of orange juice. (e) Size of family. (f) Achievement score in mathematics. (g) Merit ratings of employees. (h) Score on an introversion–extroversion scale.

8. Temperature is considered to be an example of a continuous variable. What kind of variable is formed when we measure temperature to the nearest tenth of a degree? Explain.

9. The weights of nine children were recorded to the nearest tenth of a pound, and their total weight was found to be 812 pounds. What do you think about reporting the average weight (812/9) as (a) 90.22222 pounds? (b) 90.2 pounds? (c) 90 pounds? Explain.

10. In one state, voters register as Republican, Democrat, or Independent and records of the total registration are kept. Which scale of measurement is used?

11. Instructor, assistant professor, associate professor, and professor form what kind of scale?

12. With an interval scale, is it proper to consider that an increase of 20 points is twice as much as an increase of 10 points? Explain.

13. Assume that the following series of numbers form an interval scale: 0, 1, 2, 3, . . . , 19, 20. (a) Would it still be an interval scale if we added 10 points to each score? Explain. (b) Would it still be an interval scale if we multiplied each score by 10? Explain.

14. (a) If the numbers in Problem 13 form a ratio scale, and 10 points are added to each, would we still have a ratio scale? Explain. (b) If we multiply each score by 10? Explain.

15. Dr. Jones, the history professor, administers a 30-item test the first day of class to ascertain his students' initial level of knowledge of American history. Bob got a 0, Joe a 10, Betty a 20, and Mark and Sheela each got a perfect 30. (a) In all likelihood, what kind of scale of measurement is this? Explain. (b) Can Dr. Jones conclude that Bob does not know anything about American history? Explain. (c) Is Betty's knowledge of American history twice as great as Joe's? Explain. (d) Can Dr. Jones conclude that Mark and Sheela have an equally good knowledge of American history? Explain.

CHAPTER 2

Frequency Distributions, Percentiles, and Percentile Ranks

When you have finished studying this chapter, you should be able to:

- Organize scores into a frequency distribution in table form;
- Understand the difference between apparent limits and real limits;
- Organize scores into a relative frequency distribution;
- Construct a cumulative frequency distribution and a relative cumulative frequency distribution; and
- Compute percentiles and percentile ranks.

Your history professor has just returned the midterm exams and writes the 50 scores on the board. Because her gradebook is in alphabetical order, the scores are not in numerical order, so the result looks like that shown in Table 2.1. Imagine that your score was 84. How does your score compare with others in the class?

You notice that there are some scores in the 60s and some in the 90s, but that most appear to be in the 70s and 80s. However, because you are still unsure about how your grade compares to the rest of the class, you decide to place the scores in order. The best way to do this

TABLE 2.1 *Scores for 50 Students on a History Midterm Examination*

84	82	72	70	72
80	62	96	86	68
68	87	89	85	82
87	85	84	88	89
86	86	78	70	81
70	86	88	79	69
79	61	78	75	77
90	86	78	89	81
67	91	82	73	77
80	78	76	86	83

TABLE 2.2 *Frequency Distribution of Possible Scores on History Midterm Examination (from Table 2.1), Organized by Order of Magnitude*

SCORE	FREQUENCY	SCORE	FREQUENCY
96	1	78	4
95	0	77	2
94	0	76	1
93	0	75	1
92	0	74	0
91	1	73	1
90	1	72	2
89	3	71	0
88	2	70	3
87	2	69	1
86	6	68	2
85	2	67	1
84	2	66	0
83	1	65	0
82	3	64	0
81	2	63	0
80	2	62	1
79	2	61	1

is to list all *possible* score values from the highest to the lowest. The highest score on the exam was 96 and the lowest was 61, so the recorded sequence is 96, 95, 94, . . . , 61. You now return to the unordered scores on the blackboard, take them in order of occurrence, record a tally mark for each score in the new (ordered) list, and then convert these tallies to numbers. The result, shown in Table 2.2, shows the possible scores and their frequency of occurrence. A table of observations presented in this form is called a **frequency distribution**.

Once you have organized the data in this way, you can easily see that there is one student whose grade is well above the rest, and two students who "aren't getting it." The bulk of the scores fall between 67 and 91 and your score of 84 is close to being a very good performance in this class.

2.1 Organizing Qualitative Data

Observations, or data, occur in a variety of forms. They may be numerical scores, such as weights in pounds, daily temperatures, or scores on a history midterm, or they may be qualitative data, as in the case of eye color, type of cheese, or political party. Regardless of their nature, we must

TABLE 2.3 *Frequencies of Academic Majors in a College of Science*

MAJOR	FREQUENCY
Biology	529
Chemistry	221
Physics	106
Geology	58
Mathematics	41
	Total = 955

organize and summarize observations to make sense out of them. We can make a frequency distribution for qualitative data as well as for quantitative data. This is illustrated in Table 2.3, where students enrolled in the College of Science of a local university are organized according to academic major.

frequency distribution shows the number of observations for the possible categories or score values in a set of data

The same procedures used to construct Table 2.2 were used to construct Table 2.3. However, instead of possible scores, we have different categories. Thus, we must expand our definition of a frequency distribution: A **frequency distribution** shows the number of observations for the possible categories or score values in a set of data. In Table 2.3, the order in which we put the different types of academic majors does not matter because the categories form a nominal scale. Put them in whatever order you wish; often, arranging them in order of magnitude of frequency (as in Table 2.3) will make sense. With ordinally scaled observations, you should arrange the categories in order of rank.

2.2 Grouped Scores

class intervals in a frequency distribution, a range of values that are grouped together

What should we do with numerical data when there is a wide range of values? In this case, it is often easier to understand their meaning if we combine individual scores to form groups of scores. Suppose, for example, that after constructing a frequency distribution of the history exam scores (Table 2.2), you decide to group every three consecutive scores beginning with 60, 61, and 62 and ending with 96, 97, and 98. Rather than a distribution of original scores, you would now have a distribution of score **class intervals** from 60–62 to 96–98. Column A in Table 2.4 shows how the grouped frequency distribution would appear. With this distribution, you can quickly see that most of the exam scores were between 66 (the lowest score of the interval 66 to 68) and 92 (the highest score of the interval 90 to 92).

Although grouped frequency distributions can make it easier to interpret data, some information is lost. In Column A of Table 2.4, for example, we can see that more people scored in the interval 84–86 than in any other interval. However, unless we have all the original scores to look at, we would not know whether the 10 scores in this interval were all 84s, all 85s, all 86s, or were spread throughout the interval in some way. This problem is referred to as **grouping error**. The wider the class interval width, the greater the potential for grouping error.

raw scores scores in their original form

A set of **raw scores** (scores in their original form) does not result in a unique set of grouped scores. For example, had you decided to use a class interval width of 5 instead of 3, your grouped frequency distribution might have looked like that in Column B of Table 2.4. We say "might have looked like" because even for an interval width of 5, there are several (five, in fact) possible grouped frequency distributions for the data in Table 2.2. For example, instead of starting with the interval 60–64, you might have decided to start with the interval 61–65, in which case your grouped frequency distribution would have appeared as shown in Column C of Table 2.4.

If a given set of scores can be grouped in more than one way, what is the best way so that the data are most easily understood? Fortunately, there are some widely accepted

TABLE 2.4 *Scores from Table 2.2 Converted to Grouped Frequency Distributions*

A CLASS INTERVAL WIDTH = 3		B CLASS INTERVAL WIDTH = 5		C CLASS INTERVAL WIDTH = 5	
Scores	*Frequency*	*Scores*	*Frequency*	*Scores*	*Frequency*
96–98	1	95–99	1	96–100	1
93–95	0	90–94	2	91–95	1
90–92	2	85–89	15	86–90	14
87–89	7	80–84	10	81–85	10
84–86	10	75–79	10	76–80	11
81–83	6	70–74	6	71–75	4
78–80	8	65–69	4	66–70	7
75–77	4	60–64	2	61–65	2
72–74	3		$N = 50$		$N = 50$
69–71	4				
66–68	3				
63–65	0				
60–62	2				
	$N = 50$				

conventions that can help us here. These are just guidelines and not hard-and-fast rules. Refer to Table 2.4 as you work down this list.

1. ***Be sure that your class intervals are mutually exclusive.*** That is, intervals should not overlap, so no score can belong to more than one interval.
2. ***Make all intervals the same width.*** This convention usually makes it easier to recognize the overall pattern of the data.
3. ***Make the intervals continuous throughout the distribution.*** In Column A of Table 2.4, there are no scores in the interval 63–65, but it appears anyway. To omit this interval and close the gap would create a misleading impression.
4. ***Place the interval containing the highest score value at the top.*** This convention saves the trouble of having to figure out how to read each new table when we come to it. If you can avoid it, your *highest and lowest intervals should not include any score values that are not possible.* For example, if the highest possible score on an exam was 100, it would be misleading to have 98–102 as your highest interval.
5. ***For most work, use 10 to 20 class intervals.*** In Table 2.4, the distribution in Column A meets this criterion; the distributions in Columns B and C do not. Fewer class intervals mean wider intervals and thus greater loss of accuracy. More class intervals result in greater complexity. The number of class intervals will, of course, be a function of sample size. When the number of cases is small, for example, a small number of intervals is often preferable.
6. ***Choose a convenient interval width.*** In choosing the interval width, the number of intervals desired is only one factor to consider; convenience is another. Values such as 2, 3, 4, 5, 10, 25, and 50 make both the construction and interpretation of a distribution easier, whereas values like 9, 19, or 33 should be avoided. If you are going to present your frequency distribution as a graph (Chapter 3), *selection of an odd-numbered interval width is advisable because the midpoints are whole numbers.*
7. ***When possible, make the lower score limits multiples of the interval width.*** This also makes construction and interpretation easier. The distributions in Columns A and B in Table 2.4 meet this criterion; the distribution in Column C does not.

2.3 How to Construct a Grouped Frequency Distribution

range
the difference between the lowest score and the highest score

i
symbol for interval width

f
symbol for frequency

n
symbol for total number of cases in a sample

N
symbol for total number of cases in a population

Using these guidelines, we will now organize the history midterm scores (from Table 2.1) into a grouped frequency distribution. We first find the lowest score (61) and the highest score (96). The difference between these two figures is the **range** of scores. Therefore, the range is $96 - 61 = 35$. Because there should be 10 to 20 class intervals, we must find an interval width satisfying that condition. If we divide the range by 10, we get the interval width necessary to cover the range in 10 intervals ($35/10 = 3.5$). If we use 20 intervals, the class interval width would be exactly half of this amount ($35/20 = 1.75$). We should choose an interval width equal to some *convenient* number between 1.75 and 3.5, and the only possibilities are 2 and 3. Suppose we decide to use an interval width of 3. In statistical work, interval width is symbolized by the letter ***i***; for this problem, we have selected $i = 3$.

Next, we determine the starting point of the bottom class interval. The lowest score is 61, thus the lowest interval could be 59–61, 60–62, or 61–63. We select 60–62 because 60 is a multiple of our interval width of 3. This gives us the set of class intervals shown in Table 2.5. The intervals would be the same if we had first determined the starting point of the top class interval (96–98, 95–97, or 94–96).

Next, tally the raw scores one by one against the list of class intervals. Then convert the tally to frequency (symbolized by ***f***), as shown in the last column of Table 2.5. The total number of cases in the distribution is found by summing all the values of f and is symbolized by ***n*** if the distribution is considered a sample, or by ***N*** if it is a population. These steps are summarized as follows:

Step 1: Find the lowest score and the highest score.

Step 2: Find the range by subtracting the lowest score from the highest.

Step 3: Divide the range by 10 and by 20 to determine the largest and the smallest acceptable interval widths. Choose a convenient width (i) within these limits.

Step 4: Determine the score at which the lowest (or highest) interval should begin. It should ordinarily be a multiple of the interval width.

Step 5: Record the limits of all class intervals, placing the interval containing the highest score value at the top. Make the intervals continuous and of the same width.

TABLE 2.5 *Tallies for Determining the Frequency of Scores on the History Midterm Exam (Table 2.1)*

<table>
<tr><th>SCORES</th><th>TALLY</th><th>f</th></tr>
<tr><td>96–98</td><td>|</td><td>1</td></tr>
<tr><td>93–95</td><td></td><td>0</td></tr>
<tr><td>90–92</td><td>||</td><td>2</td></tr>
<tr><td>87–89</td><td>卌 ||</td><td>7</td></tr>
<tr><td>84–86</td><td>卌 卌</td><td>10</td></tr>
<tr><td>81–83</td><td>卌 |</td><td>6</td></tr>
<tr><td>78–80</td><td>卌 |||</td><td>8</td></tr>
<tr><td>75–77</td><td>||||</td><td>4</td></tr>
<tr><td>72–74</td><td>|||</td><td>3</td></tr>
<tr><td>69–71</td><td>||||</td><td>4</td></tr>
<tr><td>66–68</td><td>|||</td><td>3</td></tr>
<tr><td>63–65</td><td></td><td>0</td></tr>
<tr><td>60–62</td><td>||</td><td>2</td></tr>
<tr><td></td><td></td><td>$N = 50$</td></tr>
</table>

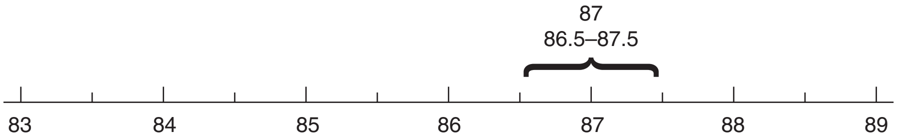


FIGURE 2.1 Real limits of a score on a history midterm.

Step 6: Using the tally system, enter the raw scores in the appropriate class intervals.

Step 7: Convert each tally to a frequency (f).

2.4 Apparent versus Real Limits

real limits of a score extend from one-half of the smallest unit of measurement below the value of the score to one-half unit above

apparent limits of a score interval extend from the smallest unit of measurement in the interval to the largest

real (exact) limits of a score interval extend from one-half the measurement unit below the lowest score to one-half the unit above the highest score

In our example, we treated the test score as if it were a discrete variable. We did not allow for the possibility of getting a fraction of a point. What would we have done if one student had scored $86\frac{1}{4}$ points? Table 2.5 shows that this score does not fall in any class interval. It lies between the class intervals 84–86 and 87–89. How, then, do we deal with continuous variables?

In general, the **real limits of a score** are considered to extend from one-half of the smallest unit of measurement below the value of the score to one-half unit above. (Age is one common exception; when a man says he is 52, he means his age is between 52.0 and 53.0.) For example, consider three possible adjacent scores on the history midterm: 86, 87, and 88. It is reasonable to think that the score of 87 represents a level of knowledge closer to 87 than that indicated by a score of 86 or 88. Because the smallest unit of measurement on the exam was 1, the real limits of the score of 87 extend from 86.5 to 87.5. We illustrate this interpretation of a score in Figure 2.1.

In Chapter 1, you learned that the process of measurement always reduces a continuous variable to a discrete one. If we record length to the nearest tenth of an inch, a measurement of 2.3 in. really represents the interval 2.25 in. to 2.35 in. If we weigh coal to the nearest 10 lb., a weight of 780 lb. represents 780 ± 5 lb., or from 775 to 785 lb.

How do we determine real limits in a grouped frequency distribution? Consider the class interval 84–86. Because a score of 84 extends down to 83.5 and a score of 86 extends up to 86.5, the interval 84–86 includes everything between 83.5 and 86.5. For this interval, 84 and 86 are only the **apparent limits**. The **real (exact) limits** are 83.5 (the *lower real limit*) and 86.5 (the *upper real limit*). Table 2.6 gives the real limits for the complete distribution of history midterm scores. Notice that the lower real limit of each class interval serves at the same time as the upper real limit of the class interval immediately below. *No score can ever fall right on a real limit because we calculate real limits by taking half the smallest unit of measurement.*

2.5 The Relative Frequency Distribution

Suppose that Catherine Candidate received 1,375 votes for mayor in the last election. Is that a large number? In a town where 1,600 people voted, it is; in fact, it is 86% of the vote. In a city of 60,000 voters, the situation is different; it amounts to a little over 2%. For some purposes, an answer to the question "How many?" may give us what we want to know. For others, the question "What proportion?" or equivalently "What percentage?" may be more relevant.

We can easily translate frequencies for each class interval of a frequency distribution to relative frequencies by converting frequencies to a proportion or percentage of the total number

TABLE 2.6 *Lower and Upper Real Limits for the Distribution in Table 2.5*

APPARENT LIMITS	REAL LIMITS	*f*
96–98	95.5–98.5	1
93–95	92.5–95.5	0
90–92	89.5–92.5	2
87–89	86.5–89.5	7
84–86	83.5–86.5	10
81–83	80.5–83.5	6
78–80	77.5–80.5	8
75–77	74.5–77.5	4
72–74	71.5–74.5	3
69–71	68.5–71.5	4
66–68	65.5–68.5	3
63–65	62.5–65.5	0
60–62	59.5–62.5	2
		$N = 50$

of cases. This gives us a relative frequency distribution. A **relative frequency distribution** shows the categories or score values and the proportion or percentage of the total number of cases that they represent. Table 2.7 shows the data of Table 2.6 expressed in this form under Class A. To obtain relative frequencies, we divide each frequency by the total number of cases in the distribution (**f/n** for a sample; **f/N** for a population). This gives the proportion of cases in the interval expressed as a decimal fraction (as a part relative to one). If you prefer percentages (parts relative to 100), move the decimal two places to the right. For example, we obtained the entry for the interval 60–62 for Class A of Table 2.7 by dividing the frequency of 2 by 50, which results in the decimal fraction .04 or 4%.

relative frequency distribution shows the proportion or percentage of the total number of observations for the possible categories or score values in a set of data

relative frequency expressed as a proportion for a sample, obtained by f/n

Relative frequencies are particularly helpful when comparing frequency distributions in which the number of cases differs. Table 2.7 shows the distributions of scores on the same history midterm examination for two classes. Class A has an enrollment of 50; Class B has an enrollment of

TABLE 2.7 *Relative Frequency Distributions of History Midterm Exam Scores*

	CLASS A			CLASS B		
APPARENT LIMITS	*f*	*rel f* (*propor.*)*	*rel f* (*%*)*	*f*	*rel f* (*propor.*)*	*rel f* (*%*)*
96–98	1	.02	2	1	.01	1
93–95	0	.00	0	1	.01	1
90–92	2	.04	4	3	.04	4
87–89	7	.14	14	3	.04	4
84–86	10	.20	20	4	.05	5
81–83	6	.12	12	7	.09	9
78–80	8	.16	16	8	.10	10
75–77	4	.08	8	9	.11	11
72–74	3	.06	6	12	.15	15
69–71	4	.08	8	6	.08	8
66–68	3	.06	6	11	.14	14
63–65	0	.00	0	7	.09	9
60–62	2	.04	4	2	.02	2
57–59				3	.04	4
54–56				3	.04	4
Totals	50	1.00	100	80	1.01	101

*When actually reporting data, one would show relative frequencies as either proportions or percentages, but not both.

80. It isn't easy to compare the two sets of raw frequencies, but it is easier to make meaningful comparisons if we convert to relative frequencies. (The percentages for the second class add to 101% rather than to exactly 100% because of rounding error.)

Relative frequencies may be inappropriate when there are few cases in the distribution. For example, it can be misleading to say that 50% of a community's auto mechanics are alcoholics when there are only two mechanics in that community.

2.6 The Cumulative Frequency Distribution

Henry earned a score of 52 on an arithmetic achievement test. How many of his class scored lower? How many scored higher? Mary earned a score of 143 on a college entrance test. Among college applicants, what percentage obtains lower scores? On the same test, what score is so low that only 10% of the applicants do worse? These are examples of questions that we can answer more easily if we organize the distribution in cumulative form.

cumulative frequency distribution shows how many cases lie below the upper real limit of each class interval

A **cumulative frequency distribution** shows how many cases lie below the upper real limit of each class interval. As an example of how we construct such a distribution, look at the frequency distribution of history exam scores shown in the first three columns of Table 2.8. We start at the bottom and record for each class interval the total frequency of cases falling *below* its upper real limit. (Remember, it is not possible to have a score exactly at the real limit—see Section 2.4.) These figures appear in the fourth column, headed *cum f.* We obtain them by adding the frequency of the given class interval to the cumulative frequency recorded for the next lower class interval. For example, we begin at the bottom by entering 2 for the two cases in the interval 60–62. This tells us that two cases fall below its upper real limit of 62.5. As we move up into the next interval, 63–65, we pick up no new cases, so the total number of cases below 65.5 remains at 2. Moving up one more interval, to 66–68, we pick up an additional three cases, giving a total of 5 ($2 + 0 + 3$) below its upper real limit of 68.5. We continue to work our way to the top by adding the frequency of each class interval to the cumulative frequency for the interval immediately below. As a check, the cumulative frequency for the uppermost class interval should equal N, the total number of cases.

If you prefer a **relative cumulative frequency distribution**, divide each cumulative frequency by n (sample) or N (population). For example, for the interval 87–89, *cum f*/N = $47/50 = .94$, with the relative cumulative frequency expressed as a proportion. These values

TABLE 2.8 *Cumulative Frequency and Relative Cumulative Frequency Distributions*

APPARENT LIMITS	REAL LIMITS	*f*	*cum f*	*cum f* (propor.)	*cum %*
96–98	95.5–98.5	1	50	1.00	100.0
93–95	92.5–95.5	0	49	.98	98.0
90–92	89.5–92.5	2	49	.98	98.0
87–89	86.5–89.5	7	47	.94	94.0
84–86	83.5–86.5	10	40	.80	80.0
81–83	80.5–83.5	6	30	.60	60.0
78–80	77.5–80.5	8	24	.48	48.0
75–77	74.5–77.5	4	16	.32	32.0
72–74	71.5–74.5	3	12	.24	24.0
69–71	68.5–71.5	4	9	.18	18.0
66–68	65.5–68.5	3	5	.10	10.0
63–65	62.5–65.5	0	2	.04	4.0
60–62	59.5–62.5	2	2	.04	4.0
		$N = 50$			

cum f
symbol for cumulative frequency

cum %
symbol for cumulative percentage

are shown in the fifth column of Table 2.8. You may convert them to percentages by multiplying each by 100, as shown in the last column. We will refer to a relative cumulative frequency distribution that uses percentages as a **cumulative percentage distribution**.

The values in the ***cum f*** column and the ***cum %*** column give the frequency and percentage of cases falling below the upper real limits of the various intervals. For instance, 40 cases (80% of the total of 50 cases) in the history class fall below 86.5, and 47 cases (94% of the total) fall below 89.5.

2.7 Percentiles and Percentile Ranks

percentile (point)
a point on the measurement scale below which a specified percentage of the cases in a distribution falls

percentile rank
the percentage of cases in a distribution that falls below a given point on the measurement scale

The percentile system is widely used to show how an individual has performed relative to a known group. It is based on the cumulative percentage distribution. A **percentile point** is a point on the measurement scale below which a specified percentage of the cases in the distribution falls. It is more commonly called a percentile (from the Latin *centrum*, meaning "hundred"). If, for example, 50% of students in the history course have midterm scores lower than 81, the 50th percentile is 81. A **percentile rank** is the percentage of cases falling below a given point on the measurement scale. In our example, the percentile rank of a score of 81 is 50.

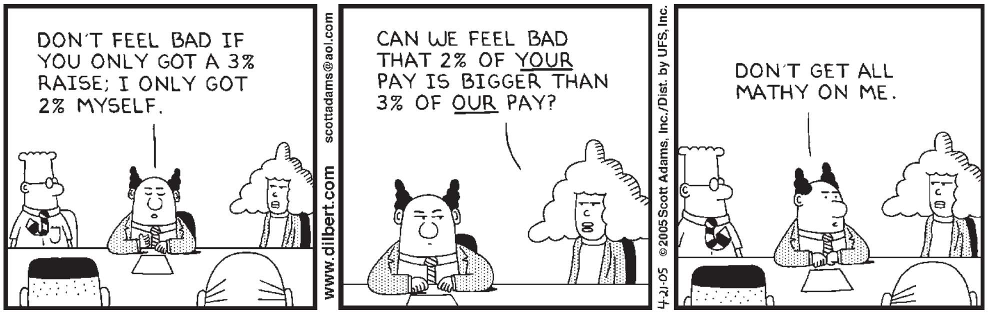
Do not confuse percentiles and percentile ranks: *Percentile ranks may take values only between 0 and 100, whereas a percentile (point) may have any value that scores may have.* For example, it is possible that the value of a percentile is 143. Suppose that Mary earned a score of 143 on a college entrance test and that 75% of applicants score lower. The 75th percentile is 143; Mary's percentile rank is 75. Standardized tests, such as the SAT (Scholastic Aptitude Test) and GRE (Graduate Record Exam), publish their results in terms of either percentiles or percentile ranks. Table 2.9 shows the percentile ranks of some scores for all those who took the GRE® in the period July 2005 through June 2008. From this table we see that a score of 600 had a percentile rank of 86 on verbal reasoning and 47 on quantitative reasoning.

TABLE 2.9 *General Test Interpretive Data for the GRE®*

Based on the performance of all examiners who tested between July 1, 2005, and June 30, 2008

	PERCENTAGES OF EXAMINEES SCORING LOWER THAN SELECTED SCALED SCORES	
SCALED SCORE	***Verbal reasoning***	***Quantitative reasoning***
800	99	94
700	97	71
600	86	48
500	62	28
400	35	13
300	8	4
Number of examinees	1,421,856	1,421,502

GRE materials selected from GRE® Guide to the Use of Scores, 2009–10. Reprinted by permission of Educational Testing Service, the copyright owner. Permission to reprint GRE materials does not constitute review or endorsement by Educational Testing Service of this publication as a whole or of any other testing information it may contain.


DON'T FEEL BAD IF YOU ONLY GOT A 3% RAISE; I ONLY GOT 2% MYSELF.
scottadams@aol.com
www.dilbert.com
CAN WE FEEL BAD THAT 2% OF YOUR PAY IS BIGGER THAN 3% OF OUR PAY?
4-21-05 © 2005 Scott Adams, Inc./Dist. by UFS, Inc.
DON'T GET ALL MATHY ON ME.

DILBERT: © United Feature Syndicate, Inc.

P
symbol for a percentile

We will use the symbol ***P*** to represent a percentile: the 50th percentile is written as P_{50}, and the 28th as P_{28}. For Mary we write: $P_{75} = 143$. The subscript indicates the percentile rank for the score of 143. In Table 2.8, we find that P_{94} (the 94th percentile point) is 89.5, the upper limit of the fourth from the top class interval. Similarly, the same table shows that $P_{60} = 83.5$ and $P_{10} = 68.5$. In the next section, you will learn how to compute percentiles from cumulative frequency distributions.

2.8 Computing Percentiles from Grouped Data

In Table 2.10, we present the data for Class B of Table 2.7 in cumulative frequency form. This is the starting point for finding percentiles. Suppose that we want to find the value of P_{25}. What is P_{25}? By definition, it is the point on the measurement scale below which 25% of the actual scores fall. There are 80 cases; because 25% of 80 is 20, P_{25} is the point below which 20 cases fall. Working up from the bottom of the distribution, we find that the 20th case will fall in the class interval 65.5–68.5. (We must be sure to think in terms of real limits.) At this point, it is not clear what score value we should assign because the point we want lies somewhere within the interval.

TABLE 2.10 *Cumulative Percentage Distribution and Percentile Calculations*

APPARENT LIMITS	REAL LIMITS	*f*	*cum f*	*cum %*	PERCENTILE CALCULATIONS
96–98	95.5–98.5	1	80	100.0	
93–95	92.5–95.5	1	79	98.8	
90–92	89.5–92.5	3	78	97.5	
87–89	86.5–89.5	3	75	93.8	
84–86	83.5–86.5	4	72	90.0	
81–83	80.5–83.5	7	68	85.0	
78–80	77.5–80.5	8	61	76.2	
75–77	74.5–77.5	9	53	66.2	② ③
72–74 ①	71.5–74.5	12	44	55.0	$P_{50} = 71.5 + (8/12 \times 3)$ $= 73.5$
69–71	68.5–71.5	6	32	40.0	
66–68	65.5–68.5	11	26	32.5	$P_{25} = 65.5 + (5/11 \times 3)$ $= 66.9$
63–65	62.5–65.5	7	15	18.8	
60–62	59.5–62.5	2	8	10.0	
57–59	56.5–59.5	3	6	7.5	
54–56	53.5–56.5	3	3	3.8	
		$N = 80$			

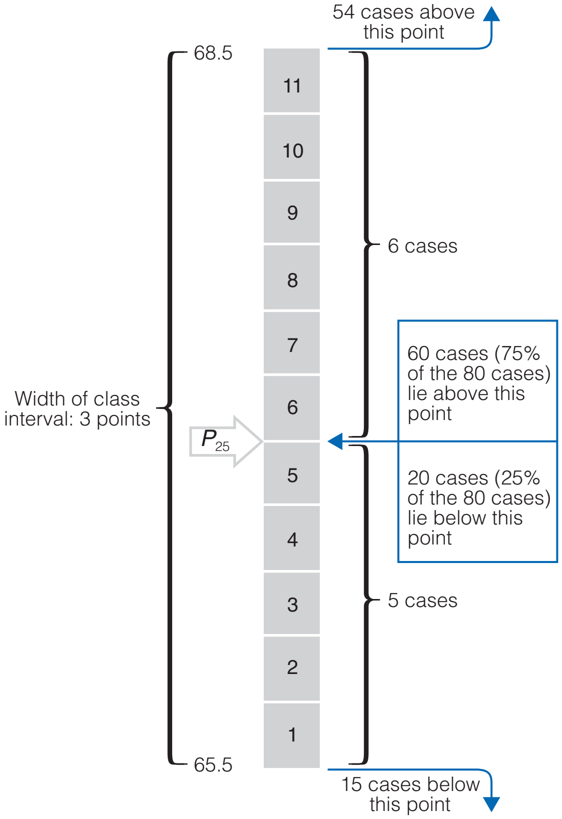

54 cases above this point
68.5
11
10
9
8
7
6
5
4
3
2
1
65.5
6 cases
60 cases (75% of the 80 cases) lie above this point
20 cases (25% of the 80 cases) lie below this point
5 cases
Width of class interval: 3 points
P_{25}
15 cases below this point

FIGURE 2.2 Location of the 25th percentile. Data from Table 2.10.

There are 11 scores in this interval. To proceed, *we will make the assumption that the scores are evenly distributed throughout the interval*. This assumption underlies the procedure called **linear interpolation**. We show the results of this assumption in Figure 2.2. Look at it closely. The value of the 25th percentile (point) will be located at a point 20 scores (cases) up from the bottom of the distribution. Because there are 15 cases below the lower real limit of the class interval concerned, we must come up 5 more to reach P_{25}. This means that we must come up 5 out of the 11 equal parts in the interval, or 5/11 of the interval's width. The width of the interval is 3 score points, and the calculation is as follows: $(5/11) \times 3 = 1.4$ points. We therefore add this quantity to the lower *real* limit of the interval.

$$P_{25} = 65.5 + 1.4 = 66.9$$

We provide a formula for this calculation in the mathematical notes at the end of the chapter, but it is more important that you follow the logic of the procedure. Once you understand the logic, it is much easier to remember than any formula. We will review and summarize the steps in calculating a percentile by determining P_{50}.

Problem

Find P_{50}.

Step 1: First objective: find the class interval in which P_{50} falls.

(a) P_{50} is the score point below which 50% of the cases fall.
(b) 50% of $N = .50 \times 80 = 40$ scores.
(c) The 40th score (from the bottom) falls in the class interval 72–74 (see ① in Table 2.10).

Step 2: Determine the number of cases between the bottom of the distribution and the lower real limit of the class interval containing the desired percentile. In this case, the number is 32.

Step 3: Determine the additional number of cases required to make up the number of scores found in 1(b). In this case, 8; 32 + 8 = 40.

Step 4: Assume the (12) scores in the class interval are evenly distributed throughout the interval.

Step 5: Find the additional distance into the interval needed to arrive at the score point: 8/12 of the interval width, or (8/12) × 3 = 2.0. This is shown at ② in Table 2.10.

Step 6: Add this additional distance to the lower real limit of the class interval to obtain the percentile

$$P_{50} = 71.5 + 2.0 = 73.5$$

See ③ in the same table.

A good way for you to check the calculation of percentiles is to count *downward* in the distribution to the appropriate point. For example, to check the calculation of P_{25}, you must come down 75% of the cases, or 60 scores. This is six cases below the upper real limit of the class interval 65.5–68.5. By the same reasoning applied previously,

$$P_{25} = 68.5 - (6/11 \times 3) = 66.9$$

A special situation arises when a percentile coincides with an empty class interval. In the data of Table 2.8, this occurs for P_4. By the definition of a percentile, P_4 could be said to be any score value between 62.5 and 65.5, inclusive. The preferred procedure is to split the difference; we therefore take P_4 as the midpoint of this interval, 64.0.

2.9 Computation of Percentile Rank

For some studies, we may want to use percentile ranks rather than percentiles. If we make the same assumption we did for the computation of a percentile, the general nature of the procedure is the same. We will work one example.

What is the percentile rank of a score of 86? We must find the percentage of cases that lie below 86.0. First, we find the number of cases below this point. We see from Table 2.10 that the score of 86.0 is located in the class interval 83.5–86.5. The interval width is 3 score points, and to reach 86.0 one must come up 2.5 of these points from the bottom of the interval (83.5 + 2.5 = 86.0). There are four cases in the interval, and we *assume them to be evenly distributed*. We must therefore come up (2.5/3) × 4 or 3.3 cases from the bottom of the interval. Because there are 68 scores below this lower limit, the point in question is 68 + 3.3 = 71.3 cases up from the bottom of the distribution. Finally, 71.3/80 = .89 or 89%. The score of 86 is at a point below which 89% of the cases fall, and therefore its percentile rank is 89. We may summarize the calculation as follows.

$$\begin{array}{l}\text{PERCENTILE RANK OF}\\ \text{A SCORE OF 86}\end{array} = 100\left[\frac{68 + \left(\frac{2.5}{3} \times 4\right)}{80}\right] = 89$$

To be certain you understand the procedure, verify the following calculation of the percentile rank of a score of 61.

$$\text{PERCENTILE RANK OF A SCORE OF 61} = 100\left[\frac{6 + \left(\frac{1.5}{3} \times 2\right)}{80}\right] = 9$$

For further practice, verify that the percentile rank of a score of 71 is 39, and that the percentile rank of a score of 90 is 94.

2.10 Summary

If we have a large number of observations, it is easier to understand their meaning if we order and group them. A frequency distribution shows the number of observations for the possible categories or score values in a set of data. Although grouping (putting scores into class intervals) is often preferred, there is a cost—we no longer know exactly where the scores are in a particular class interval. For this reason, class intervals should not be too wide. A good compromise for most purposes is to use between 10 and 20 intervals.

We can construct frequency distributions to show frequency (how many?) or relative frequency (what proportion or percentage of the whole?). Relative frequency distributions are usually best for comparing two or more distributions containing different numbers of cases. The cumulative frequency distribution gives the number of scores below the upper real limit of each score interval.

A percentile rank is the percentage of cases falling below a given point on the measurement scale. Thus, it allows us to see how an individual has performed relative to the entire group. A percentile is a point along the measurement scale below which a specified percentage of the cases in the distribution falls. Testing services usually use this system when publishing results of standardized tests.

Mathematical Notes

In this chapter, we begin a series of mathematical notes. They are not intended to provide rigorous derivations of statistical procedures. We offer them to provide additional insight into statistical relationships to the student who wants to know more about the subject. *The notes are part of an "enriched curriculum." However, the text is designed to stand on its own without these notes.*

Note 2.1 Computation of Percentile Points (*Ref.*: Section 2.8)

To calculate a percentile point, you may use the following general formula.

$$P = LL + (i)\left(\frac{cum\,f\,\text{percentile} - cum\,f\,\text{below}}{f}\right)$$

where:

LL = lower real limit of class interval containing the percentile

i = width of the class interval

$cum\,f$ percentile = number of scores lying below the percentile

$cum\,f$ below = number of scores lying below LL

f = frequency of scores in the interval containing the percentile

Note 2.2 Computation of Percentile Rank (*Ref.*: Section 2.9)

$$\text{Percentile Rank} = 100\left(\frac{\mathit{cum}\, f\,\text{below} + \left(\frac{X - LL}{i}\right)(f)}{N}\right)$$

where: $\mathit{cum}\, f$ below = number of scores below lower real limit of class interval containing X

X = score

LL = lower real limit of class interval containing X

i = width of the class interval

f = frequency of scores in the interval containing X

N = number of scores in the distribution

Key Terms, Concepts, and Symbols

frequency distribution (p. 19)
class intervals (p. 19)
grouping error (p. 19)
raw scores (p. 19)
range (p. 21)
i (p. 21)
f (p. 21)

n (p. 21)
N (p. 21)
real limits (p. 22)
apparent limits (p. 22)
relative frequency distribution (p. 23)
cumulative frequency distribution (p. 24)

relative cumulative frequency distribution (p. 24)
cum f (p. 25)
cum % (p. 25)
percentile (point) (p. 25)
percentile rank (p. 25)
P (p. 26)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Students in introductory psychology were asked to name their favorite foreign food. Their responses were as follows:

Chinese	Mexican	Italian	Greek
Mexican	Italian	Italian	Chinese
Indian	Chinese	Chinese	Greek
Mexican	Indian	Japanese	Chinese
Mexican	Chinese	Chinese	Mexican
Italian	Chinese	Mexican	Chinese

(a) Organize the results into a frequency distribution. (b) On what measurement scale are these observations? (c) Does it matter in what order we arrange the categories?

2. Students in introductory statistics were asked about their academic standings. They gave the following responses:

freshman	junior	sophomore	junior
sophomore	graduate	sophomore	junior
sophomore	junior	sophomore	senior
senior	senior	junior	sophomore
sophomore	senior	senior	freshman
junior	graduate	sophomore	senior

(a) Organize the results into a frequency distribution. (b) On what measurement scale are these observations? (c) How does this frequency distribution differ from that in Problem 1? Hint: Does it matter in what order we arrange the categories?

3. Write the real limits for the following scores: (a) A score of 52; measurement is to the nearest digit. (b) 800 yd.; measurement is to the nearest 100 yd. (c) 460 lb.; measurement is to the nearest 10 lb. (d) .6 in.; measurement is to the nearest .1 in. (e) .47 sec.; measurement is to the nearest .01 sec.

4. For each of the following intervals, give the interval width, the real limits, the apparent limits of the next higher interval, and the real limits of the next higher interval. (a) 5–9, where the scores were rounded to the nearest whole number. (b) 40–49, where the scores were rounded to the nearest whole number. (c) 2.0–2.4, where the scores were rounded to the nearest tenth. (d) 60–70, where scores are accurate to the nearest ten. (e) 1.75–1.99, where scores were rounded to the nearest hundredth.

5. In making a frequency distribution, you set the real limits for the lowest two intervals at 29.5–39.5 and 39.5–49.5. A friend objects, saying, "The intervals overlap, because both contain 39.5." (a) Reply to your friend's objection. (b) Give the apparent limits and the interval width for both intervals.

6. For each set of class intervals, identify any shortcomings.

(a)	(b)	(c)
50 and up	20–25	5–9
44–49	14–19	10–14
38–43	8–13	14–19
26–31	0–7	20–24

7. Given below are the highest and the lowest score for different samples. Each sample is to be grouped into class intervals. For each, give the range, your choice of width for the class intervals, the apparent limits for the lowest interval, and the apparent limits for the highest interval: (a) 75, 36; (b) 117, 54; (c) 171, 27; (d) $+21$, -22; (e) 3.47, 1.13; (f) 821, 287.

8. Data 2A gives the scores earned by college students on an anthropology test:

DATA 2A

44	35	20	40	38	52	29	36	38	38
38	38	41	35	42	50	31	43	30	41
32	47	43	41	47	32	38	29	23	48
41	51	48	49	37	26	34	48	35	41
38	47	41	33	39	48	38	20	59	37
29	44	29	33	35	58	41	38	26	29
32	54	24	38	38	56	56	48	34	35
26	26	38	37	57	24	44	62	29	41

(a) Construct a grouped frequency distribution with 20–22 as the apparent limits for the lowest class interval. The interval width should thus be 3. (b) Construct another frequency distribution for the same scores, again with a width of 3 for the class intervals, but begin with 18–20 as the lowest interval. (c) List the lower and upper real limits for each interval. (d) Compare the two distributions. Are they generally similar in the impression they give of the raw scores?

9. When you group scores following the guidelines in Section 2.2, does it matter whether you start with the highest scores or the lowest scores?

10. Convert the following proportions to percentages, preserving the accuracy in the proportion: (a) .73; (b) .09; (c) .666; (d) .07; (e) .008.

11. Convert the following percentages to proportions, preserving the accuracy in the percentage: (a) 37%; (b) 56.3%; (c) 4.2%; (d) 9.21%; (e) .4%; (f) .85%; (g) .02%.

12. (a) To the frequency distribution that you constructed for Problem 8a, add a column of relative frequencies expressed as proportions. (b) To the frequency distribution that you constructed for Problem 8b, add a column of relative frequencies expressed as percentages.

13. Two hundred volunteers from a large section of the introductory psychology course participate in an experiment testing the effect of mood on learning. Some spend 10 minutes thinking of pleasant things to put them into a good mood; the others spend the time thinking of unpleasant things and go into a bad mood. All subjects then devote 30 minutes to studying an unfamiliar, sorrowful folk tale from an African culture, after which they take a test on it. Data 2B gives the frequency distributions of scores on this test. (The higher the number, the better the learning of the folk tale.)

(a) Present the distributions in terms of relative frequency, using percentages. (b) Suppose you wish to compare the two conditions of the experiment with regard to the results for the range 130–139 inclusive. Which provide the more meaningful comparison: the original frequencies or the relative frequencies? Explain.

Data 2B

SCORE	GOOD MOOD	BAD MOOD
155–159		1
150–154	2	2
145–149	4	7
140–144	7	12
135–139	12	10
130–134	14	7
125–129	25	4
120–124	23	3
115–119	18	0
110–114	20	2
105–109	12	1
100–104	8	0
95–99	3	1
90–94	2	
	$n = 150$	$n = 50$

14. To the tables that you constructed for Problems 8a and 13a, (a) add a column giving cumulative frequency, and then (b) add a column giving cumulative percentage.

15. This problem requires no computation. For the data displayed in Table 2.8, find: (a) the 24th percentile; (b) P_{60}; (c) the 10th percentile; (d) P_{94}; (e) the percentile rank of the value 62.5; (f) the percentage of cases with a score less than 77.5; (g) the percentile rank of the value 95.5; (h) the percentage of cases with a score below 59.5.

16. (a) Can a percentile have the value of 517? Explain. (b) Can a percentile have the value -5.8? Explain. (c) Can a percentile rank have the value 517? Explain. (d) Can a percentile rank have the value -5.8? Explain.

17. From the cumulative distribution of the Good Mood scores in Data 2B (Problems 13 and 14), find: (a) the percentile rank of 139.5; (b) the percentile rank of 126; (c) the 25th percentile; (d) the 50th percentile.

18. As part of a preventative medicine program, a mobile medical unit obtains a variety of measures for each of the 30 employees of a local bank. The following are the results for diastolic blood pressure:

Data 2C

81	91	89	81	79	82
70	92	80	64	73	86
87	72	74	75	90	85
83	82	79	82	78	96
77	85	83	87	88	80

(a) Construct a grouped frequency distribution. (b) Construct a cumulative frequency distribution. (c) Construct a cumulative percentage distribution.

19. Dr. Smart's first exam in introductory statistics was on frequency distributions. The scores are given in Data 2D. (The highest possible score was 100.)

Data 2D

52	84	93	78	75
71	99	81	86	81
65	70	72	71	91
87	82	77	66	63
90	58	89	60	79
77	72	83	87	87
83	79	55	97	74
71	86	75	83	63
82	70	90	95	92
75	85	83	71	88

(a) Construct a grouped frequency distribution. (b) Construct a cumulative frequency distribution. (c) Construct a cumulative percentage distribution.

CHAPTER 3

Graphic Representation of Frequency Distributions

When you have finished studying this chapter, you should be able to:

- Know how to construct and interpret frequency distributions, relative frequency distributions, and cumulative frequency distributions in graph form; more specifically;
- Know how to construct a bar diagram and pie chart for qualitative data;
- Know how to construct a histogram and frequency polygon for quantitative data;
- Know how to construct a cumulative percentage curve;
- Understand the criteria for choosing between a histogram and frequency polygon; and
- Understand the factors that affect the shape of graphs, and identify some shapes that are common in the behavioral sciences.

Should you construct a frequency distribution as a table or a graph? A graph is based entirely on the tabled data and therefore can tell us nothing more than we can learn by inspecting the table. However, graphic representation often makes it easier for us to see pertinent features of the data. A study that compared the use of tables and graphs in science concluded that graphs are highly preferred because "of their relative readability, ease of comprehension, combinability, and overall rhetorical effectiveness" (Smith et al., 2002).

There are many different kinds of graphs. Books are available that describe graphic procedures at length (Cleveland, 1995; Cleveland & McGill, 1985). We will consider only frequency distributions, which can be represented in graphic form in three main ways: the *histogram* and its variant, the *bar diagram*; the *frequency polygon*; and the *cumulative percentage curve.* These graphs are illustrated in Figures 3.1, 3.6, 3.2, and 3.9, respectively. When you first inspect them, the four graphs may appear to be very different from one another, but they have much in common. Therefore, before describing each one in detail, we will discuss some general procedures in graphing frequency distributions.

3.1 Basic Procedures

abscissa the horizontal or ***X*** **axis**

ordinate the vertical or ***Y*** **axis**

zero point on a graph the intersection of the abscissa and ordinate

Graphed frequency distributions generally have two axes: horizontal and vertical. The horizontal axis is called the **abscissa** or ***X*** **axis**, and the vertical axis the **ordinate**, or ***Y*** **axis**. *It is customary to represent scores or categories (for qualitative data) along the horizontal axis and frequency (or some function of frequency) along the vertical axis*. According to mathematical practice, the intersection of the two axes represents the **zero point on a graph**. If it does not, the reader should be warned on the graph. In Figure 3.1, for example, a break in the horizontal axis indicates that a portion of the scale is missing.

The appearance of a graph depends on the choice of scale used to represent position along the two axes. Because similar distributions should appear similar when graphed, *it is customary to make the height of the distribution about three-quarters of the width.* We measure width and height from the span of the graphed data rather than from the borders of the graph. It is important that a graph be large enough so that others can easily read the values (both along the horizontal and vertical axes and within the graph itself). *The graph of any frequency distribution should also have a succinct and informative title as well as labels on both axes.*

By now you can see that some procedures are basically arbitrary and governed by conventions. These conventions are useful because they result in representations that match our expectations. Then we do not have to approach each new graph as a problem-solving exercise before absorbing its meaning. Keep these general procedures in mind as we look at the different types of graphed frequency distributions.

3.2 The Histogram

histogram a graph that consists of a series of rectangles, the heights of which represent frequency or relative frequency

Suppose that after constructing a grouped frequency distribution of your history exam scores (see Table 3.1), your history professor wants to display the results as a graph. There are two basic ways in which she could do this. One of them is the histogram shown in Figure 3.1. The **histogram** consists of a series of rectangles, each of which represents the scores in one of the class intervals of the distribution. *The two vertical boundaries of a rectangle coincide with the real limits of the particular interval, whereas its height depends on the frequency of scores for that interval.* We can use either raw frequencies or relative frequencies (proportions or percentages) with a histogram. Changing from raw to relative frequencies requires only that we relabel the vertical axis. The steps in constructing a histogram are summarized as follows (see Table 3.1 and Figure 3.1 for examples):

Step 1: Construct a frequency distribution in table form as described in Section 2.3 of the previous chapter.

Step 2: Before drawing your axes (graph paper is advised), decide on a suitable scale for the horizontal axis (scores) and then determine the number of squares (on the graph

TABLE 3.1 *Results from 50 Students on the History Class Midterm Examination*

A: RAW SCORES

84	82	72	70	72
80	62	96	86	68
68	87	89	85	82
87	85	84	88	89
86	86	78	70	81
70	86	88	79	69
79	61	78	75	77
90	86	78	89	81
67	91	82	73	77
80	78	76	86	83

B: FREQUENCY AND CUMULATIVE FREQUENCY DISTRIBUTIONS

Apparent limits	*Real limits*	*f*	*rel f (%)*	*cum f*	*cum %*
96–98	95.5–98.5	1	2	50	100.0
93–95	92.5–95.5	0	0	49	98.0
90–92	89.5–92.5	2	4	49	98.0
87–89	86.5–89.5	7	14	47	94.0
84–86	83.5–86.5	10	20	40	80.0
81–83	80.5–83.5	6	12	30	60.0
78–80	77.5–80.5	8	16	24	48.0
75–77	74.5–77.5	4	8	16	32.0
72–74	71.5–74.5	3	6	12	24.0
69–71	68.5–71.5	4	8	9	18.0
66–68	65.5–68.5	3	6	5	10.0
63–65	62.5–65.5	0	0	2	4.0
60–62	59.5–62.5	2	4	2	4.0
	Totals =	50	100%		

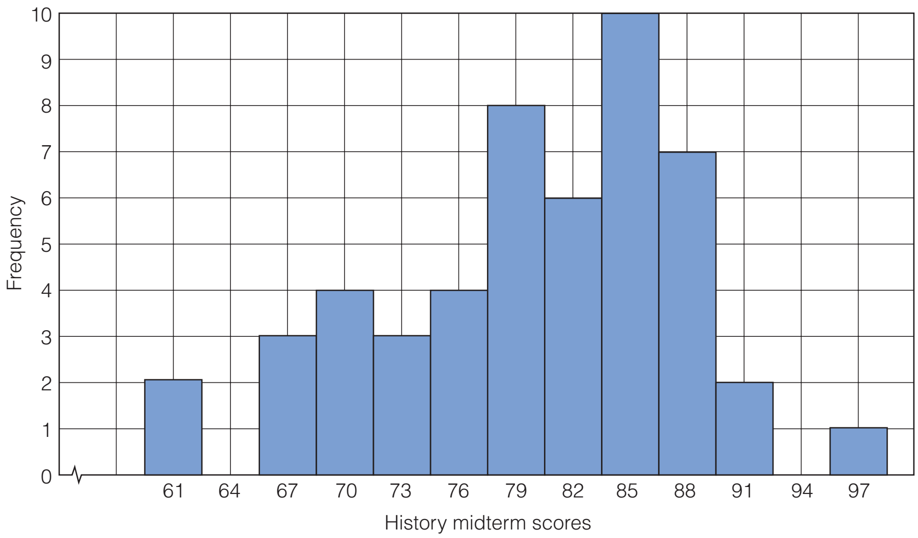


FIGURE 3.1 Histogram of grouped history exam scores from Table 3.1.

paper) required for the width of the graph. Multiply this number by .75 to find the approximate number of squares to use for the graph's height (frequency).

Step 3: Draw bars of equal width for each class interval. The height of a bar corresponds to the frequency or relative frequency in that particular interval. *There should be no gaps between the bars except for intervals in which there are no scores* (such as 63–65 and 93–95 in Figure 3.1). The edge of a bar represents both the upper real limit for one interval and the lower real limit for the next higher interval.

Step 4: Identify your class intervals along the horizontal axis by using either real limits or interval midpoints. Do not use both because it will result in too many numbers along the abscissa and may confuse the reader. *If you use real limits, place them under the edge of each bar* (see Figure 3.10). *If you use interval midpoints, place them under the middle of each bar* (see Figure 3.1). You find the **interval midpoint** by locating the point halfway between the apparent limits (or, alternatively, the real limits) of the interval. Remember, odd-numbered interval widths have midpoints with whole numbers.

interval midpoint the point halfway between the apparent (or real) limits of an interval

Step 5: Label your axes and give your histogram a title.

3.3 The Frequency Polygon

Another way in which your history professor could construct a graph of the distribution in Table 3.1 is with a frequency polygon (from the Greek meaning "many angles"). As you can see in Figure 3.2, a **frequency polygon** is a series of connected dots above the midpoint of each possible class interval. Each dot is at a height equal to the frequency or relative frequency of scores in that interval. The steps in constructing a frequency polygon are summarized as follows:

frequency polygon a graph that consists of a series of connected dots above the midpoint of each possible class interval (height of the dots corresponds to frequency or relative frequency)

Step 1: Construct a frequency distribution in table form.

Step 2: Decide on a suitable scale for the horizontal and vertical axes.

Step 3: Label your class interval *midpoints* along the horizontal axis.

Step 4: Place a dot above the midpoint of each class interval at a height equal to the frequency or relative frequency of the scores in that interval.

Step 5: Connect the dots with straight lines.

Step 6: Label your axes and give your polygon a title.

Notice that in Figure 3.2 the polygon touches the horizontal axis at 64 and 94, the midpoints of the intervals 63–65 and 93–95, respectively. This indicates that there were zero scores in those intervals. Ordinarily, we bring the polygon down to the horizontal axis at both ends as well. To do so, identify the two class intervals falling immediately outside those end class intervals containing scores. The midpoints of these intervals, plotted at zero frequency, are then connected to the graph. We did this for the interval 57–59 because scores in this interval were possible, though not obtained. However, what do we do when scores in the next adjacent class

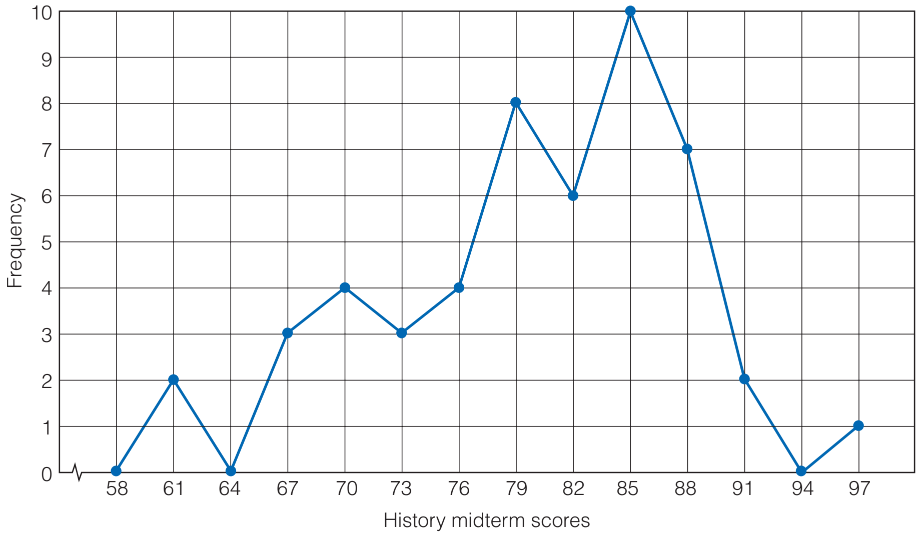


Figure 3.2 Frequency polygon of grouped history exam scores from Table 3.1.

interval are not possible? There are problems with any solution, but we suggest the best thing to do in a case like this is to leave the dot "dangling."[1] We have done this in Figure 3.2 for the interval 96–98 because scores greater than 100 were not possible on the history exam and the next adjacent interval is 99–101.

3.4 Choosing between a Histogram and a Polygon

Both the histogram and the polygon are used for graphing quantitative data on an interval or ratio scale. You can see the similarities between the two graphs by superimposing a histogram and polygon of the same set of data, as we have done for the history exam scores in Figure 3.3. Nevertheless, there are occasions when one may be preferred over the other.

A histogram is often used when graphing an ungrouped frequency distribution of a discrete variable (or data treated as a discrete variable). For example, Figure 3.4 shows a histogram of the results of the mathematics portion of the ACT (American College Testing Program) national profile for high school graduates between 2006 and 2008. The ACT is used by over 1,000 colleges and universities to assess their freshmen's level of knowledge. Scores are given in whole numbers from 1 to 36. In this case, the histogram helps convey to the reader that there are no possible scores other than those reported on the horizontal axis. The general public seems to find a histogram a little easier to understand than a polygon, and hence it may be a good choice for communicating with them.

A histogram also has some merit when displaying relative frequency. The total area in a histogram represents 100% of the scores, and thus the area in the bars of a histogram is directly representative of relative frequency. That is, *the area in any rectangle is the same fraction of the total area of the histogram as the frequency of that class interval is of the total number of cases in the distribution*.

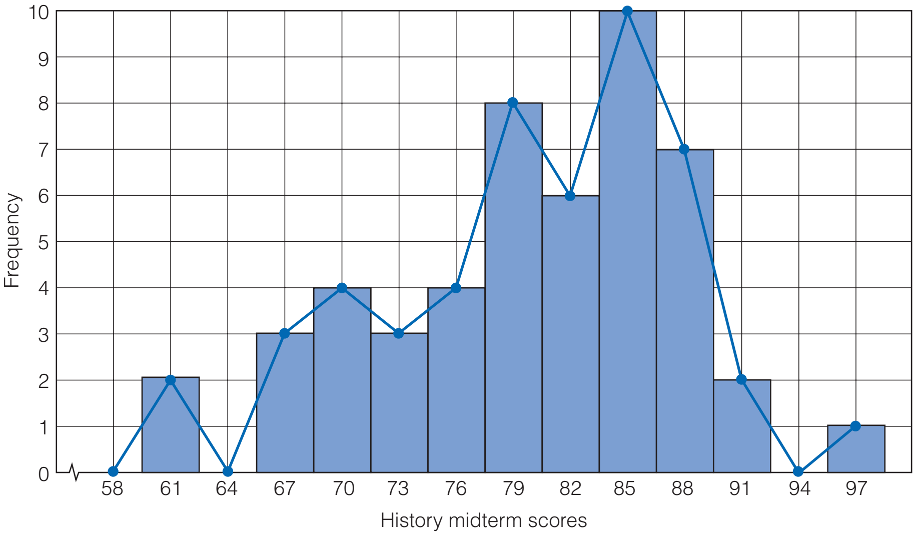


FIGURE 3.3 Frequency polygon of grouped history exam scores superimposed on a histogram of the same scores.

---

[1] Bringing the polygon down to 0 at the midpoint of the next adjacent interval may mislead someone looking at the graph to think that scores in that interval were possible. Bringing the polygon straight down implies there were two frequencies for that interval, including $f = 0$.

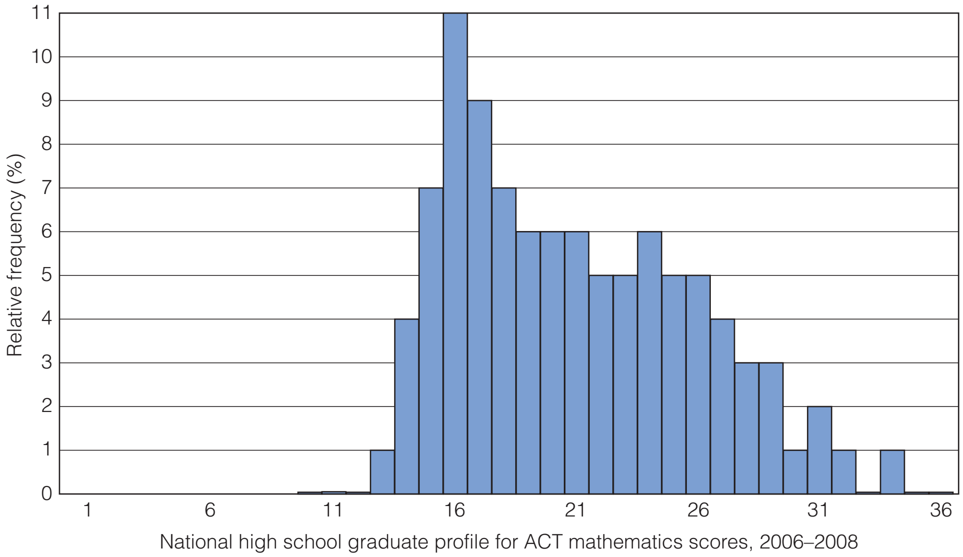


FIGURE 3.4 Histograms are often used when graphing ungrouped distributions of a discrete variable. (*Source:* Taken from ACT Class Profile Report, high school graduates 2006, 2007, and 2008. Copyright by ACT, Inc. Reproduced with permission.)

A polygon is often preferred for grouped frequency distributions. Representing frequencies by rectangular bars suggests that the scores are evenly distributed within each class interval and that the borders of the intervals are points of decided change. If a definite trend of increasing or decreasing frequency exists over a span of several consecutive class intervals, the frequency polygon will show this trend more directly. This is because the direction of each straight line in the polygon is determined by the frequencies in two adjacent class intervals, whereas the horizontal top of each rectangle in a histogram is responsive only to what occurs in one class interval.

Frequency polygons are particularly helpful when comparing two or more distributions. Figure 3.5 shows such a comparison, using the two sets of data in Table 2.7. When distributions are based on different numbers of cases, as in this example, we can equalize that difference by using relative frequencies rather than raw frequencies. Figure 3.5 makes it easy to see that the bulk of the scores for Distribution A appear higher on the scale and that the range of performance is a little less than for Distribution B.

3.5 The Bar Diagram and the Pie Chart

What should we do when graphing a distribution of qualitative data? In this case, there are two widely used possibilities. Figure 3.6 shows a bar diagram of the distribution of new psychology Ph.D.s in major employment settings. The **bar diagram** is very similar to the histogram and is constructed in the same manner except that *space appears between the rectangles, thus properly suggesting the essential discontinuity of the several categories.* However, within categories, subcategories may be displayed as adjacent bars. For example, in Figure 3.7, the level of education achieved by people in 1968 and 2000 (different categories) is shown for both men and women (subcategories). Because qualitative categories on a nominal scale of measurement have no necessary order, you may arrange them in any order (e.g., in order of their frequencies as in Figure 3.6, if desired). However, for ordinal scales of measurement, you should arrange the categories in order of rank (e.g., freshman, sophomore, junior, senior).

bar diagram used for qualitative data, a graph that is similar to a histogram, except that space appears between the rectangles

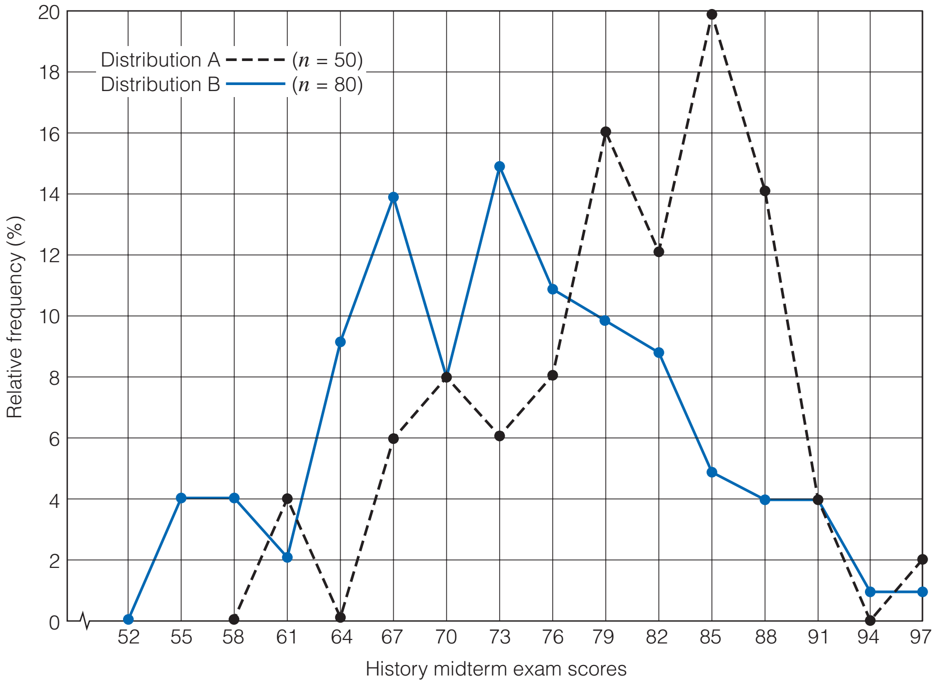


FIGURE 3.5 Relative frequency polygons used to compare the two score distributions in Table 2.7.

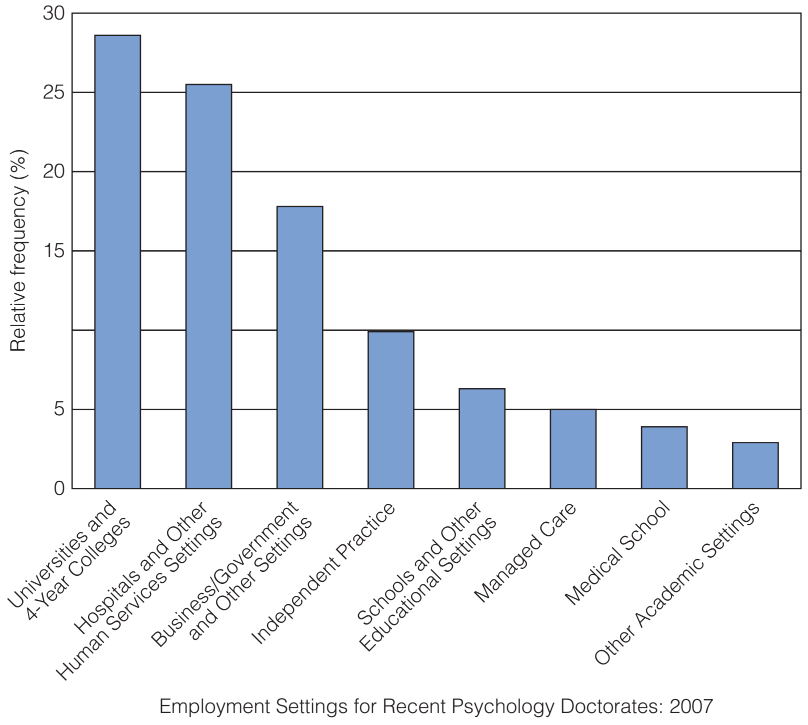


FIGURE 3.6 Primary employment settings of new American Ph.D.s in psychology, health services provider subfields. (*Source:* Copyright © 2009 by the American Psychological Association. Reprinted [or adapted] with permission from 2007 Doctorate Employment Survey. APA Center for Workforce Studies. June 2009.)

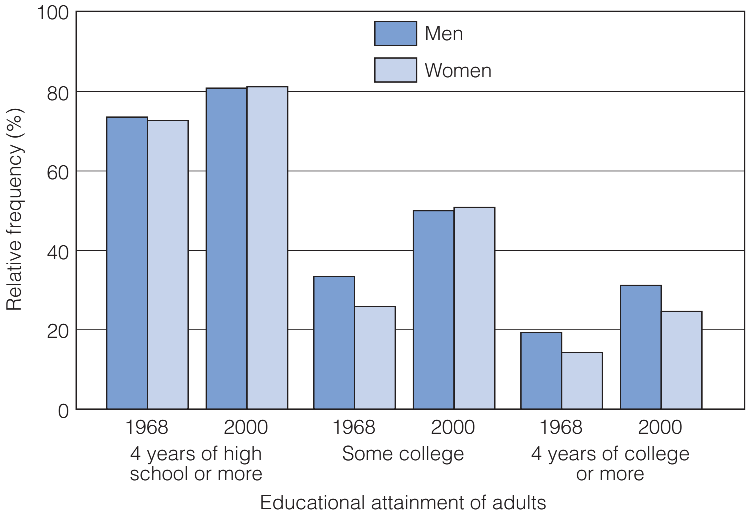
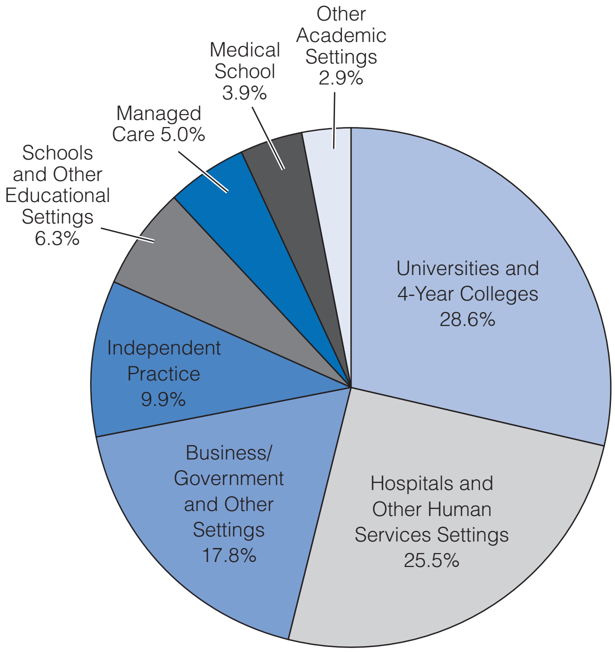


FIGURE 3.7 Example of a bar diagram for categorical data with subcategories. (*Source:* Census Bureau.)

pie chart used for qualitative data, area in any piece of the pie shows the relative frequency of a category

Another frequently used graph for qualitative data is the **pie chart**. Figure 3.8 shows a pie chart of the same data as in Figure 3.6. Unlike the bar diagram, in which results can be expressed as either raw frequencies or relative frequencies, *pie charts always use relative frequencies.* The area in any piece of the pie is the same fraction of the pie as the frequency of that category is of the total number of cases in the distribution.



Employment Settings for Recent Psychology Doctorates: 2007

FIGURE 3.8 Example of a pie chart, showing the primary employment settings of new American Ph.D.s in psychology. (*Source:* Copyright © 2009 by the American Psychological Association. Reproduced [or adapted] with permission from 2007 Doctorate Employment Survey. APA Center for Workforce Studies. June 2009.)

3.6 The Cumulative Percentage Curve

Both the cumulative frequency distribution and the cumulative percentage distribution can be presented in graphic form. Because the cumulative percentage curve is widely used in the behavioral sciences, we will illustrate only this curve.

Recall that a cumulative percentage indicates the percentage of scores that lie below the upper real limit of the associated class interval (see Section 2.6). Therefore, in constructing a **cumulative percentage curve**, the cumulative percentage is *plotted at the upper real limit* of the class interval. Note that this procedure differs from that for the frequency polygon, for which we plot class interval frequencies at the midpoint of the interval. As an example, let us refer again to the results of the midterm history exam (see Table 3.1). In Figure 3.9, the upper real limit for each class interval is labeled on the horizontal axis, and cumulative percentage appears on the vertical axis. We then place a dot over the upper real limit of each interval corresponding to the cumulative percentage for the interval. For example, the cumulative percentage for the interval 84–86 is 80.0%; therefore, we place a dot at the intersection of 86.5 (the upper real limit of the interval) and 80.0%. We then connect the dots with straight lines and *bring the curve down to zero at the lower end* at the upper real limit of the next adjacent class interval (in which there were zero scores).

cumulative percentage curve
a graph that consists of a series of connected dots above the upper real limits of each possible class interval (height of the dots corresponds to cumulative percentage)

A cumulative percentage curve never has a negative slope (that is, it never comes down). For class intervals in which there are zero scores (such as 63–65 and 93–95), the cumulative percentage curve remains horizontal. The curve shows a small rise in intervals with relatively few scores and a sharp rise in intervals with many scores. You can see this by comparing the frequency histogram or polygon of the history exam scores (Figures 3.1 and 3.2) with the cumulative percentage curve in Figure 3.9. Like the distribution of history exam scores, many distributions have most of the cases in the middle portion of the distribution. This results in a cumulative percentage curve with an S-shaped figure called an **ogive** (ō′jīv) **curve**.

ogive curve
a curve with an S shape

We may determine percentiles or percentile ranks from the cumulative percentage curve with nearly the same results obtained from the computational procedures outlined in the previous chapter. Connecting the dots on the cumulative curve with straight lines is the graphic equivalent of assuming (as we did when computing) that scores are evenly spread throughout

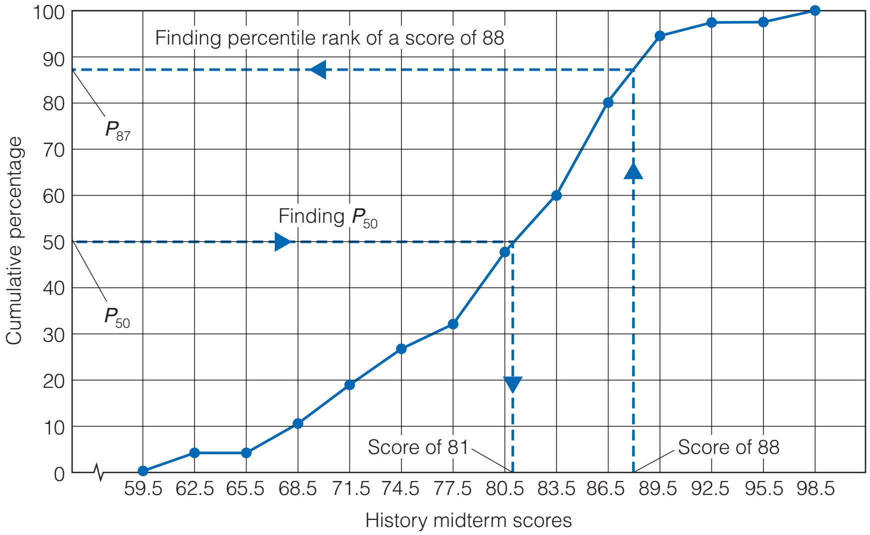


FIGURE 3.9 Cumulative percentage curve of the grouped history exam scores from Table 3.1.

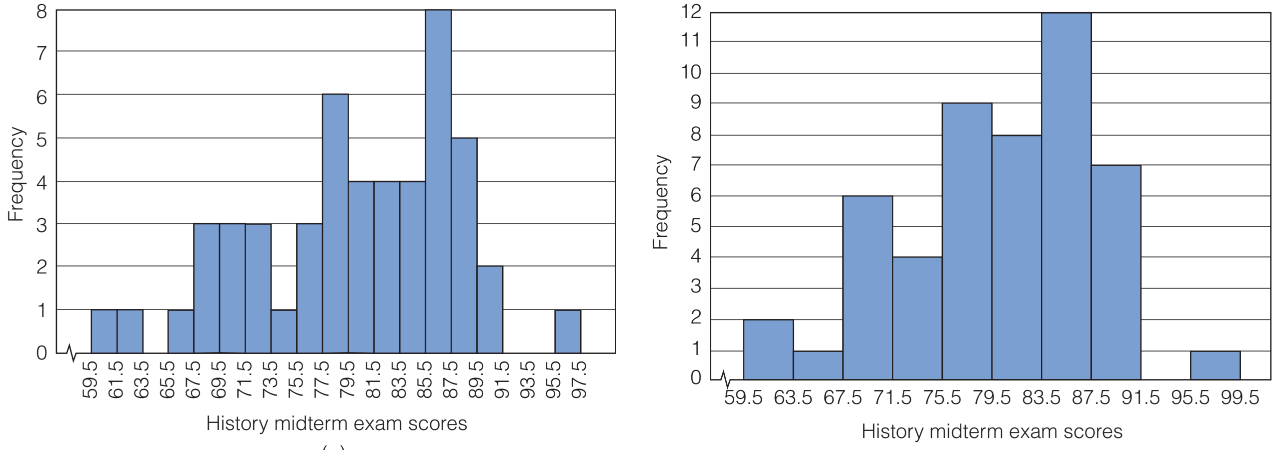
the interval. For the history exam scores illustrated in Figure 3.9, for example, we can find the percentile rank of a score of 88 by locating it on the horizontal axis and then moving vertically upward until we intersect the curve. From there we read horizontally to the vertical axis and find the percentile rank to be 87. The process may be reversed as well. To find P_{50}, follow the second set of dashed lines from the vertical axis to the intersection with the curve, and then descend to the horizontal axis. Thus, $P_{50} = 81$.

3.7 Factors Affecting the Shape of Graphs

There is no such thing as *the* graph of a given set of data. The same set of raw scores may be grouped in different ways, and the grouping will affect the graph of the distribution. For example, the two graphs in Figure 3.10 again show the results of the history exam scores, but grouped with interval widths of 2 and 4, respectively. Compare the appearance of these two graphs with each other and with Figure 3.1, where $i = 3$. Note how the appearance changes with interval width. All three graphs are properly drawn, and none can be considered to be the "correct" graph for the history exam scores.

Even more important is the matter of relative scale. Frequencies and scores are like apples and cows; no principle specifies what is an equivalent number of each. Consequently, the decision about relative scale is arbitrary, and the resulting graph can be squat or slender depending on the choice. Figure 3.11 illustrates two different graphs of the data shown in Figure 3.1. The large difference in appearance created by differences in scale is the reason for the convention that the height of the figure should be about three-quarters of the width.

Even if the convention regarding height and width is followed, the same data can appear very different when graphed, depending on the scale of measurement used for frequency. Suppose, for example, that two of your classmates are instructed by your history professor to construct a graph of the number of men and women who scored in the top half of the class on the history exam. Both students construct bar diagrams with the same relative scale for



(*a*) (*b*)

FIGURE 3.10 Histograms of grouped history exam scores from Table 3.1 with (*a*) $i = 2$ and (*b*) $i = 4$. Notice that for even-numbered interval widths, both the real limits and the midpoints (not shown) are not whole numbers.

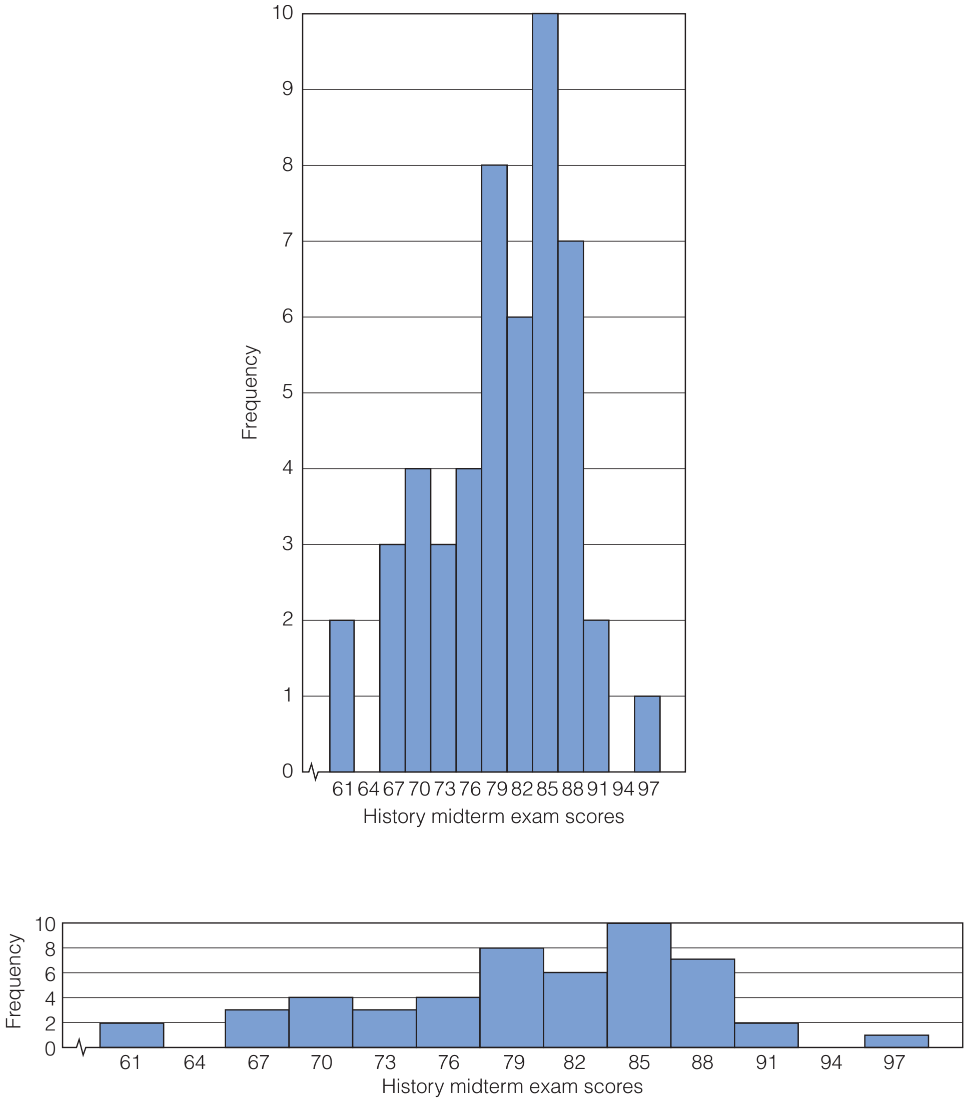


FIGURE 3.11 Histograms of grouped history exam scores drawn with different relative scales than that shown in Figure 3.1. In all cases the width of the class intervals is 3.

height and width (see Figure 3.12), but guess which one wishes to convey that the women were far superior to the men? You will sometimes see graphs with a break in the vertical axis as in Figure 3.12. Never do this! *Frequency on the vertical axis should always be continuous from zero*. When we put a break in the axis, we lose the proportional relationship among class interval frequencies. In Figure 3.12*b,* for example, it incorrectly appears as if women did over twice as well as men. Figure 3.12*a* shows the correct proportional relationship.

3.8 Shape of Frequency Distributions

What is the pattern of the distribution of scores over the range of possible values? Are most of the scores in the middle, at one end, or clustered in two distinct locations? Certain shapes of frequency distributions occur with enough regularity in statistical work that they have

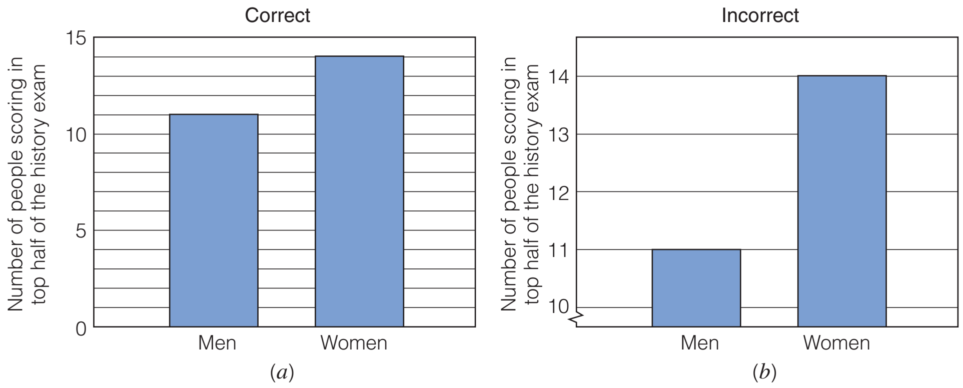


FIGURE 3.12 Two bar diagrams showing the same results using different scales for frequency. (*a*) The left graph shows the correct proportional relationship between men and women. (*b*) In the right graph, putting a break in the vertical axis results in an incorrect proportional relationship.

skewed distribution a distribution in which one tail slants to the left (negatively skewed) or to the right (positively skewed)

names. The names effectively summarize the general characteristics of the distribution. We illustrate several of them in Figure 3.13. The **J-shaped distribution** in Figure 3.13*a* could result from plotting the speeds at which automobiles go through an intersection where a stop sign is present. J-shaped distributions can also have tails that go to the left (from which they take their name). Figures 3.13*b* and 3.13*c* both show **skewed** (asymmetrical) **distributions**. Figure 3.13*b* is *positively skewed*, or skewed to the right; it might result from a test that is too difficult for most of the group taking it. Figure 3.13*c* is *negatively skewed*, or skewed to the left; it might result from a test that is too easy for most of the group. Figure 3.13*d* is an example of a **rectangular distribution**, where there are an equal number of cases in all class intervals. Figure 3.13*e* shows a **bimodal distribution**; it could result from measuring strength of grip in a group that contained both men and women.

kurtosis refers to the degree of peakedness of a graphed distribution

Figure 3.13*f* shows a **bell-shaped distribution**. A specific type of bell-shaped distribution, called the *normal curve*, is of great importance in statistical inference. Most of Chapter 6 will be devoted to the normal distribution. **Kurtosis** refers to the degree of peakedness of a graphed distribution. If a distribution is flatter than the normal curve, we refer to it as *platykurtic*. It is called *leptokurtic* if more peaked, and *mesokurtic* if it has the same peakedness as a normal distribution.

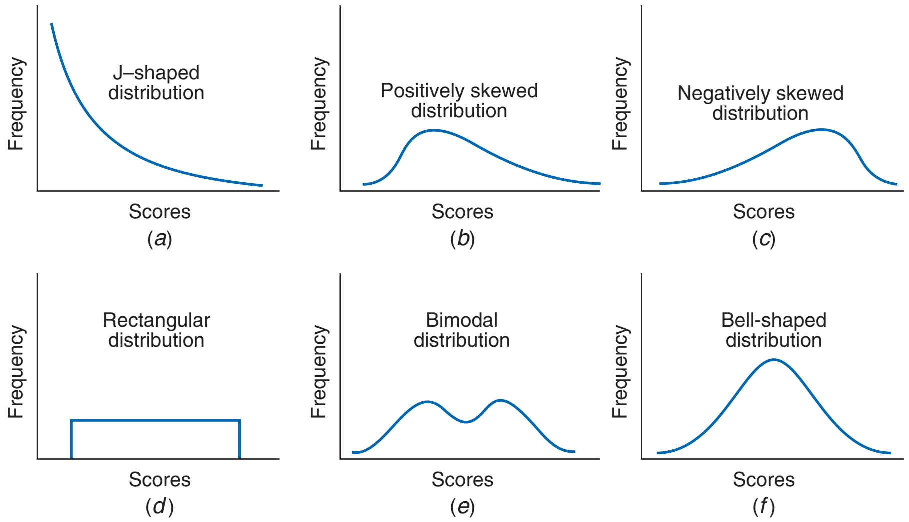


FIGURE 3.13 Shapes of some distributions that occur in statistical work.

BASIC RESEARCH
APPLIED SCIENCE
TECHNOLOGY
MANUFACTURING
ADVERTISING

Source: © Sidney Harris, ScienceCartoonsPlus.com

3.9 Summary

Graphs often make it easier to see certain characteristics and trends in a set of data. The histogram and the frequency polygon are two important ways of representing a frequency distribution of quantitative data. In the histogram, frequency is represented by the height of adjacent bars erected at the real limits of the class intervals. In the frequency polygon, frequency is indicated by dots plotted above the midpoint of each interval. These dots are then connected by straight lines. Frequency or relative frequency can be plotted with either graph. A special advantage of the histogram is the exact relationship that exists between area in a bar (or bars) and the frequency of cases in the distribution. It is often preferable for discrete variables, particularly when scores are ungrouped. On the other hand, the frequency polygon better suggests the continuity of a variable and is often preferable for a grouped distribution or for comparing distributions.

Bar diagrams and pie charts are used for qualitative data. The gaps between the bars of a bar diagram emphasize the discontinuity of categorical data.

Cumulative percentage is graphed by plotting dots above the upper real limit for each class interval. The straight lines between the dots allow us to quickly find percentile ranks and percentiles (when the graph is carefully drawn).

Conventions for constructing graphs help ensure that people working with the same data will construct similar graphs. Nevertheless, different pictures can result from grouping scores in different ways and from using different scales on the horizontal or vertical axes. Thus, there is no such thing as *the* graph of a set of scores.

Key Terms, Concepts, and Symbols

abscissa (p. 34)
X axis (p. 34)
ordinate (p. 34)
Y axis (p. 34)
zero point on a graph (p. 34)
histogram (p. 34)
interval midpoint (p. 36)
frequency polygon (p. 36)
bar diagram (p. 38)

pie chart (p. 40)
cumulative percentage curve (p. 41)
ogive curve (p. 41)
J-shaped distribution (p. 44)
skewed distribution (p. 44)
rectangular distribution (p. 44)
bimodal distribution (p. 44)
bell-shaped distribution (p. 44)
kurtosis (p. 44)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In making a histogram of the data for Class B in Table 2.7, you would draw a series of bars. For each of the following class intervals, state over what point on the abscissa you would place the left side of the bar and over what point you would place the right side of the bar: (a) 57–59, (b) 66–68, (c) 75–77, (d) 84–86, (e) 93–95.

2. (a) Construct a histogram from the data for Class B in Table 2.7. (b) Construct a relative frequency histogram (expressed as percentages) from the same data.

3. In making a frequency polygon of the data for Class B in Table 2.7, you would place dots over various locations on the abscissa. For each of the following class intervals, state over what point the dot would go: (a) 60–62, (b) 69–71, (c) 78–80, (d) 87–89, (e) 96–98.

4. (a) Construct a frequency polygon from the data for Class B in Table 2.7. (b) Construct a relative frequency polygon (using percentages) from the same data.

5. On a 30-point political science exam, scores ranged from 2 to 30. When constructing a frequency polygon with $i = 3$, discuss the problems with bringing the polygon down to the horizontal axis from the midpoint of the intervals 1–3 and 28–30. What should you do in a case like this?

6. (a) Construct a frequency histogram from the data for the Good Mood condition in Data 2B (Problem 13) in Chapter 2. (b) Construct a frequency polygon from the same data.

7. Construct (a) a histogram, (b) a frequency polygon, and (c) a cumulative frequency distribution in graph form for Data 2D (Problem 19) in Chapter 2. (d) Why might it have been better to have chosen an odd-numbered interval width?

8. Construct a bar diagram for the data in Problem 1 of Chapter 2.

9. (a) Construct a bar diagram for the data in Problem 2 of Chapter 2. (b) How does this bar diagram differ from the one you drew for Problem 8?

10. In a university, male psychology majors are distributed as follows: 24 freshmen, 61 sophomores, 109 juniors, 104 seniors, and 92 graduate students. Comparable figures for females are 74, 58, 99, 53, and 67, respectively. Construct a bar diagram suitable for comparing the sexes. What conclusions do the data suggest?

11. From the data in Problem 10, construct a pie chart for males.

12. For the data in Problem 1 of Chapter 2, construct a relative frequency distribution in graph form using percentages.

DATA 3A: *Reaction Time in Milliseconds*

TIME	SIMPLE STIMULUS f	COMPLEX STIMULUS f
300–319	1	3
280–299	1	6
260–279	2	10
240–259	4	18
220–239	3	25
200–219	6	35
180–199	11	28
160–179	12	16
140–159	8	7
120–139	2	2
	$n = 50$	$n = 150$

13. Plot relative frequency polygons for the two sets of scores in Data 3A on the same graph. Compare the two distributions and record your conclusions.

14. For the reaction times to complex stimuli in Data 3A: (a) Tabulate the cumulative frequencies. (b) Convert the cumulative frequencies to cumulative percentages and tabulate them. (c) Plot the cumulative percentage curve. (d) Find graphically P_{20} and P_{60}. Show with dashed lines how you found P_{20}. (e) Find graphically the percentile rank of scores of 195 and 245. Show with dashed lines how you found the percentile rank for 195.

15. Repeat parts (a) through (d) of Problem 14 for the Good Mood condition in Data 2B in Chapter 2. Find the percentile rank of a score of 137.

16. After you have completed Problem 15, find graphically the 20th percentile point.

17. Figure 3.5 shows the frequency polygons for the two distributions given in Table 2.7. (a) Construct cumulative percentage curves for these distributions, plotting them on the same graph for comparison. (b) How does the higher level of performance in Distribution A show up in these curves? (c) How does the greater spread of Distribution B show up in these curves? (d) Study your cumulative curves in comparison with the noncumulative ones in Figure 3.5. Which type of representation appears smoother? (e) For what kind of question is each type of curve best suited to provide information?

18. The cumulative percentage curve tends toward an S shape when the frequency distribution has more scores in the center than elsewhere. Draw an approximation to the cumulative percentage curve that would result if the noncumulative distribution had a shape like that pictured in (a) Figure 3.13*b;* (b) Figure 3.13*c;* (c) Figure 3.13*d;* (d) Figure 3.13*e.*

CHAPTER 4

Central Tendency

When you have finished studying this chapter, you should be able to:

- Understand what is meant by "measure of central tendency";
- Calculate the mode, median, and mean for ungrouped and grouped distributions;
- Appreciate the unique properties of the mode, median, and the mean;
- Describe where each measure of central tendency falls in symmetrical and skewed distributions; and
- Understand how each of the measures of central tendency is affected by adding a constant to each score, subtracting a constant from each score, or multiplying or dividing each score by a constant.

measure of central tendency a single summary figure that describes the central location of an entire distribution of observations

Where, on the scale of possible scores, is a point that best represents the overall level (high, medium, low) of the set of scores? Measures of this type are called **measures of central tendency**. Their purpose is to provide a single summary figure that best describes the central location of an entire distribution of observations. Two distributions that differ only in central tendency are shown in Figure 4.1.

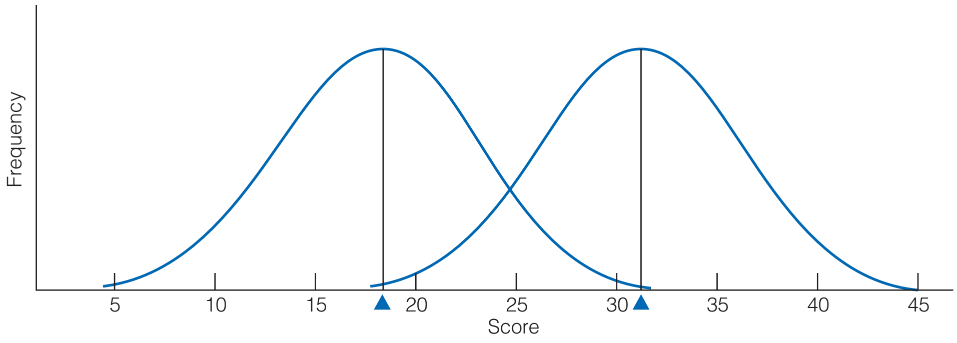


FIGURE 4.1 Distributions that differ only with respect to central tendency.

A measure of central tendency is helpful in comparing the performance of a group with that of a standard reference group. It also helps simplify comparison of two or more groups tested under different conditions. For example, we may be interested in whether video games can improve mental or physical functioning in the elderly. In one study, psychologists administered an IQ test to 11 volunteers in their 60s and 70s, then gave them 30 minutes of play on a video game twice a week for 2 months, and then tested them again. The average IQ was 101.8 before the 2-month intervention and 108.3 afterward (Drew & Waters, 1985). We would need to know more than this to conclude that video games improve mental functioning in the elderly, but as a first step, a comparison of the group averages suggests that it might.

There are several measures of central tendency. We will consider only the three most commonly used in the behavioral sciences: *mode*, *median*, and *mean*. Each of them has unique properties that make them useful in different situations.

4.1 The Mode

mode
the score that appears with the greatest frequency

Mo
symbol for the mode

A common meaning of *mode* is "fashionable," and its meaning in statistics is similar. In ungrouped distributions, the **mode** is the score that occurs with the greatest frequency. For example, in Table 2.2 the mode is 86. In grouped data, it is calculated as the midpoint of the class interval that contains the greatest number of scores. Thus, if we did not know the raw scores, we would calculate the mode for the grouped scores in Column A of Table 2.4 as 85. The symbol for the mode is ***Mo***.

4.2 The Median

median
the value that divides the distribution into halves; another name for P_{50}.

The **median** of a distribution is the point along the scale of possible scores that has half the scores below it and half the scores above it. You can see that the median is another name for the 50th percentile (P_{50}). Its symbol is ***Mdn***.

For raw scores, think of the median as the middle score of a distribution based on score frequency. To find the median, *we first put the scores in rank order* from lowest to highest. (We must include scores of zero as well.) If n (or N) is odd, the median will be the score that has an equal number of scores below and above it. For example, for the following scores:

0, 7, 8, 11, 15, 16, 20

Mdn
symbol for the median

the median is 11. When there is an even number of scores, there is no middle score, so the median is the point halfway between the two scores that are in the middle position. For example, for the group of scores:

12, 14, 15, 18, 19, 20

the median is $15 + (18 - 15)/2 = 16.5$. The only problem occurs when there are repeating scores. For example, what is the median for the following set of scores?

5, 7, 8, 8, 8, 8

The median is halfway between the first 8 and second 8. In cases like this, some people prefer to simply call the median 8. However, determining the median precisely requires interpolation. In this example, 8 represents all scores between the lower real limit of 7.5 and the upper real limit of 8.5. Because the median comes after the first of the four 8s, or one-quarter of the way through the interval width of 1.0, we calculate the median by interpolation as $7.5 + (1/4 \times 1.0) = 7.75$.

We gave the procedures for calculating P_{50} for grouped data in Section 2.8 and provided a formula to find a percentile point in Mathematical Note 2.1. The formula for finding P_{50} is as follows:

COMPUTATIONAL
FORMULA FOR
THE MEDIAN

$$Mdn = P_{50} = LL + (i)\left(\frac{.5n - cum\,f\,\text{below}}{f}\right) \tag{4.1}$$

where:
LL = lower real limit of class interval containing P_{50}
i = width of the class interval
$.5n$ = half the cases (i.e., number of scores lying below the median)
$cum\,f$ below = number of scores lying below LL
f = frequency of scores in the interval containing the median

Thus, to calculate the median for the grouped data in Table 2.10,

$$\begin{aligned} P_{50} &= 71.5 + (3)\left(\frac{40 - 32}{12}\right) \\ &= 71.5 + (3)(8/12) \\ &= 71.5 + 2 \\ &= 73.5 \end{aligned}$$

This is the same answer we arrived at in Section 2.8 without an explicit formula. The formula only expresses mathematically the steps we outlined in that section. Review that section and be sure you understand the concepts before using the formula.

4.3 The Mean

The **mean** is the sum of all the scores in a distribution divided by the total number of scores. Many people call this measure the average, but we will avoid this term because it is sometimes used indiscriminately for any measure of central tendency.

mean
the sum of all the scores divided by the total number of scores

$\boldsymbol{X}$
symbol that serves as a collective term to specify a particular set of scores

Σ
capital Greek letter sigma, read "the sum of"; gives instructions to sum whatever follows

$\overline{\boldsymbol{X}}$
symbol for the mean of a sample of scores (X)

$\boldsymbol{\mu_X}$
Greek letter mu with subscript; symbol for the mean of a population of scores (X)

$$\boldsymbol{\mu_X = \frac{\Sigma X}{N}}$$

formula for the mean of a population

$$\boldsymbol{\overline{X} = \frac{\Sigma X}{n}}$$

formula for the mean of a sample

You will need to learn some symbolism to express the mean mathematically. We will use the capital letter $\boldsymbol{X}$ as a collective term to specify a particular set of scores. (Be sure to use capital letters; lowercase letters are used in a different way.) We identify an individual score in the distribution by a subscript, such as X_1 (the first score), X_8 (the eighth score), and so forth. Remember that n stands for the number of scores in a sample and N for the number in a population. The last score in a sample can therefore be symbolized X_n; in a population, it would be X_N. Thus, a set of scores in a sample may be represented like this:

$$X\!: X_1, X_2, X_3, \ldots, X_n$$

When two groups of scores are involved, we will use the letter Y to symbolize the second set.

When scores are to be summed, we use the capital Greek letter sigma, Σ, to indicate this. It should be read "the sum of" whatever follows. It is not called "sigma" because the lowercase sigma (σ) has a different meaning in statistics (as you will see in the next chapter). Here is an example:

$$\sum X = X_1 + X_2 + X_3 + \cdots + X_n$$

The distinction between the mean of a sample and that of a population is important in inferential statistics, so there are two symbols for the mean. The mean of a sample is represented by $\boldsymbol{\overline{X}}$, read as "X bar," or $\overline{Y}$ for a second sample. The mean of a population is symbolized by $\boldsymbol{\mu}$ (the Greek letter *mu,* pronounced "mew"). The defining formulas are:

DEFINITIONAL (RAW SCORE) FORMULAS FOR THE MEAN

$$\mu_X = \frac{\sum X}{N} \qquad \text{(mean of a population)} \qquad (4.2a)$$

$$\overline{X} = \frac{\sum X}{n} \qquad \text{(mean of a sample)} \qquad (4.2b)$$

This distinction between sample and population is not crucial at a purely descriptive level, so to avoid unnecessary complication, we will use symbolism appropriate to samples whenever possible in these early chapters.

As an example, let us calculate the mean for the three scores 31.0, 21.0, and 42.0:

$$X\!: 31.0, 21.0, 42.0$$

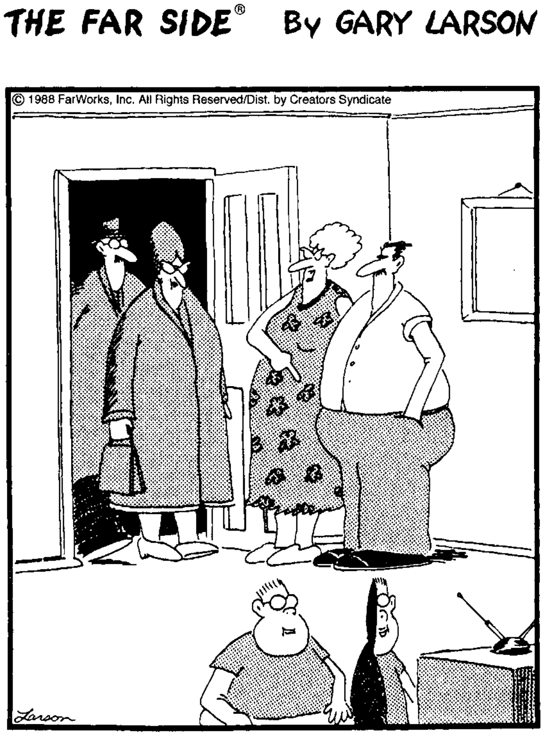
$$\sum X = 31.0 + 21.0 + 42.0 = 94.0$$

Therefore, the mean for this distribution is:

$$\overline{X} = \frac{\sum X}{n} = \frac{94.0}{3} = 31.3$$

4.4 Properties of the Mode

The mode is the only measure that can be used for data that have the characteristics of a nominal scale. For example, no other measure of central tendency is appropriate for a distribution of eye color. For quantitative variables, you need to be aware of some concerns about the use of


THE FAR SIDE® By GARY LARSON
© 1988 FarWorks, Inc. All Rights Reserved/Dist. by Creators Syndicate
Larson

"Bob and Ruth! Come on in. ... Have you met Russell and Bill, our 1.5 children?"

The Far Side® by Gary Larson © 1988 FarWorks, Inc. All Rights Reserved. The Far Side® and the Larson® signature are registered trademarks of FarWorks, Inc. Used with permission.

the mode as a measure of central tendency. When quantitative data are grouped, the mode may be strongly affected by the width and location of class intervals. Look, for example, at Table 2.4. In addition, *there may be more than one mode for a particular set of scores.* (This, however, may be very useful information.) In a rectangular distribution, the ultimate is reached: Every score shares the honor! For these reasons, the mean or the median is often preferred with numerical data.

4.5 Properties of the Mean

Unlike the other measures of central tendency, *the mean is responsive to the exact position of each score in the distribution.* Inspect the basic formula $\Sigma X/n$. It shows that increasing or decreasing the value of any score changes ΣX and thus also changes the value of the mean.

The mean is the balance point of a distribution, to use a mechanical analogy. If we imagine a seesaw consisting of a fulcrum (balance point), a board, and the scores of a distribution spread along the board like bricks (one brick for each score), the mean corresponds to the position of the fulcrum when the system is in balance. You can see this in Figure 4.2. As with the ordinary seesaw, if one brick (score) is shifted, the balance point will also change.

$(X - \overline{X})$ a **deviation score;** shows by how much a score differs from the mean

There is an algebraic way of stating that the mean is the balance point for a distribution: $\Sigma(X - \overline{X}) = 0$.[1] This says that if we express the scores in terms of the amount by which they deviate from their mean, taking into account the negative and positive deviations (scores below

---

[1] $\Sigma(X - \overline{X})$ means: $(X_1 - \overline{X}) + (X_2 - \overline{X}) + \cdots + (X_n - \overline{X})$. For example, for the three scores 3, 4, 5, $\Sigma(X - \overline{X}) = (3 - 4) + (4 - 4) + (5 - 4) = 0$.

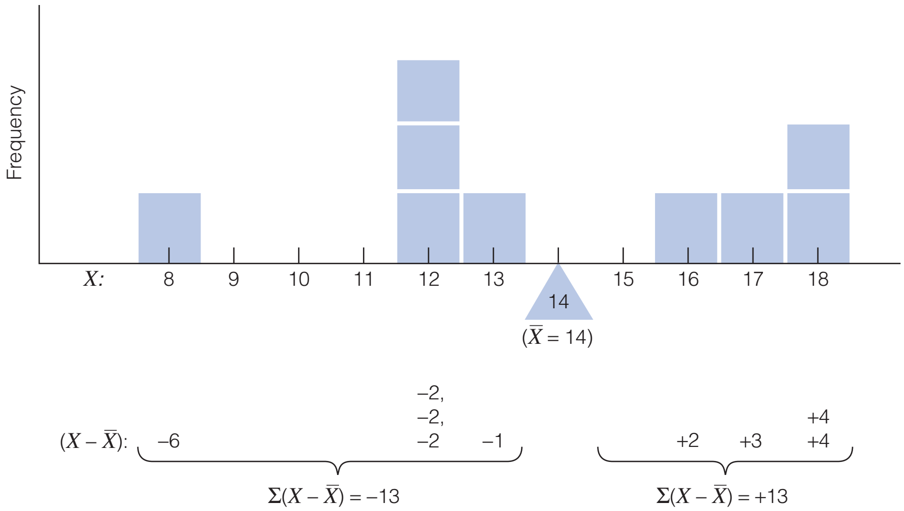


FIGURE 4.2 The mean as the balance point of a distribution.

the mean deviate negatively from the mean), their sum is zero. To put it another way, *the sum of the negative deviations from the mean exactly equals the sum of the positive deviations.* Figure 4.2 shows that $\Sigma(X - \overline{X}) = 0$ for the data given. We provide a general proof of this proposition in Note 4.1 at the end of the chapter.

When a measure of central tendency should reflect the total of the scores, the mean is the best choice because it is the only measure based on this quantity. If a track coach wants to know whether the four best quarter-milers have improved as a group, the mean will best suit the purpose because the coach has an underlying interest in the total time as an indicator of how the runners would do as a relay team. Similarly, insurance companies express life expectancy as a mean because it is most closely related to their primary concerns, which are total dollars of income from policyholders and total payoffs to survivors.

Point of Controversy

Is It Permissible to Calculate the Mean for Tests in the Behavioral Sciences?

Colleges should be free to pay salaries to their athletes.
Strongly disagree −3 −2 −1 0 +1 +2 +3 Strongly agree

Questionnaires with items like this are common. To indicate their attitudes, respondents circle numbers.

How should we summarize a sample of such numbers? Most researchers would do a little arithmetic—addition and division—and find the mean. But critics (e.g., Siegel & Castellan, 1988) would say, "Not so fast. Those operations are permissible only if the scores derive from measurement of the most sophisticated kinds. Are you sure your measurement is good enough to justify calculating a mean?"

The most sophisticated measurements employ a ratio scale. As Section 1.5 reported, a ratio scale is one whose zero point represents an absence of the quantity that the scale measures. But "0" on the scale above certainly does not indicate the absence of an attitude; it merely signifies an attitude that is neither favorable nor unfavorable—perhaps ambivalence or indecision.

Interval-level measurement would still justify calculating means, according to the critics. Does the scale above fall at least at the interval level? It's doubtful. Consider two attitudes, one represented by -3 and the other by -2. The difference between those attitudes in favorability to salaries for college athletes is probably not the same as the difference in favorability between attitudes represented by, say, $+1$ and $+2$. So a onepoint difference between scores does not necessarily signify an equal difference in attitudes all along the scale.

Does the scale then fall only at the ordinal level of measurement? If so, the seven numbers along the scale would indicate only a rank ordering from the least favorable attitude to the most favorable. But there is probably more information in the numbers than that. A two-point difference between scores probably signifies a greater difference in favorability than a one-point difference.

Measurement on this scale is therefore likely to lie somewhere between the ordinal and the interval levels of sophistication. The same is true with many other measuring instruments used in the behavioral sciences—with inventories of moods like anger and elation, with assessments of personality traits like extraversion and conscientiousness, with tests of aptitudes and achievements. They do not yield scores that carry as much intrinsic meaning as, say, temperatures and weights, but they tell us more than ranks do.

So are we or are we not justified in calculating the mean to summarize a sample of such scores? Some say yes and others no (Davison & Sharma, 1988, 1990; Maxwell & Delaney, 1985; Townsend & Ashby, 1984). By a strict interpretation of the theory of measurement presented in Section 1.6, no; we should use the median. (And in doing inferential work on such data, according to the strict interpretation, we should avoid the popular *t* and *F* tests described in later chapters.) One defense of the theory of Section 1.6 thunders, "Performing sophisticated analyses that are appropriate for one scale of measurement . . . on data that reflect a less structured scale . . . yields nonsense. The nonsense might be interpreted, but it will be nonsense nonetheless" (Stine, 1989, p. 154). In an extreme example, some researchers have even calculated means and performed advanced statistical tests on data that fall only at the nominal scale of measurement (see Whitley, 1992).

However, most researchers are not so strict, at least for data that lie somewhere between the ordinal and interval scales or higher (e.g., Glass & Hopkins, 1996; Kirk, 1990). They reject the counsel of those who would sharply limit the "permissible statistics" for the scores common in the behavioral sciences.

Some authorities even reject the theory that distinguishes the four levels of measurement (e.g., Adams, 1966; Rozeboom, 1966). In the physical sciences, it may make sense to posit some real quantity, such as heat or mass, that researchers hope to capture in numbers, but in the behavioral sciences the phenomena of interest are typically hypothetical. Attitudes, moods, traits, and abilities do not literally exist. Instead of the four levels of measurement, James Terwilliger (1978) of the University of Minnesota argued that we should distinguish three ways to index phenomena: classification systems, ranking techniques, and measurement procedures. Instead of rigidly limiting the algebraic operations on researchers' and teachers' scores, he would welcome any operations that yield lawful relationships and accurate predictions.

One of the most important characteristics of the mean is its stability from sample to sample. Suppose we draw a series of samples from a population of scores. Each sample is the same size, and each is a random selection from the available population of scores. Because chance affects which scores get into the samples, the means of these samples will vary. The medians of the samples will also vary, and so will the modes. However, if a distribution is bell-shaped (resembling the curve shown in Figure 3.13*f*), the means of the successive samples will vary the least among themselves. That is, *under these circumstances, the mean is most resistant to chance sampling variation*. This resistance makes the mean a valuable indicator of what is true about the population with regard to central tendency. We will discuss this virtue further when we get to statistical inference.

4.6 Properties of the Median

The median responds to how many scores lie below (or above) it, but not to how far away the scores may be. Thus, *the median is less sensitive than the mean to the presence of a few extreme scores* (called **outliers**). Consider, for example, the money earned by the top 200 professionals on the men's PGA (golf) tour in 2009 (see Table 4.1). The mean earning was $1,253,638, but the median was $837,886. The money earned by the best player, Tiger Woods, was over $4 million greater than that of the second-ranked player, Steve Stricker, and Stricker's earnings were over $1 million greater than the third-ranked player. The earnings of these two players strongly affected the total, and hence the mean, but their values did not affect the median. If Tiger Woods had earned $5 million more, the median would still have been the same. In this case, the median is closer than the mean to the aggregate of all the salaries. Therefore, in *distributions that are strongly asymmetrical (skewed) or have a few very deviant scores, the median may be the better choice for measuring the central tendency* if we wish to represent the bulk of the scores and not give undue weight to the relatively few deviant ones.

outlier
a very extreme score in a distribution (see Chapter 5)

In behavioral studies, there are occasions when a researcher cannot record the exact values of scores at the upper end of a distribution. For example, suppose that social psychologists studying behavior in emergencies wish to record reaction times. During a 5-minute ride between two stops on a subway line, a confederate of the experimenters pretends to collapse. How long does it take before a bystander comes over to help? On most trials, somebody does come to see what's wrong, and the reaction time ranges from a few seconds to a few minutes. But what about those few trials in which nobody moves until the next stop and the confederate has to get up to avoid being trampled by boarding passengers? (It's happened, as Piliavin, Rodin, and Piliavin, 1969, observed.) The score for such a trial should be greater than the time it took to get to the next stop, but we do not know how much greater. Distributions like this one are called open-ended. In **open-ended distributions**, we cannot calculate the mean without making assumptions, but we can find the median.

open-ended distribution
a distribution for which exact scores cannot be recorded at one end of the distribution

TABLE 4.1 *Money Earned by the Top 200 Players on the 2009 PGA Tour*

RANK	PLAYER	MONEY ($)
1	Tiger Woods	10,508,163
2	Steve Stricker	6,332,636
3	Phil Mickelson	5,332,755
4	Zach Johnson	4,714,813
5	Kenny Perry	4,445,562
.	.	.
.	.	.
96	James Nitties	931,532
97	Kevin Stadler	925,514
98	Michael Letzig	896,478
99	Lee Janzen	871,187
100	Ted Purdy	838,707
.	.	.
.	.	.
196	Dudley Hart	158,399
197	Greg Kraft	156,686
198	Kirk Triplett	155,480
199	John Huston	135,476
200	Michael Sim	130,188

4.7 Measures of Central Tendency in Symmetrical and Asymmetrical Distributions

skewed distribution a distribution in which one tail slants to the left (**negatively skewed**) or to the right (**positively skewed**)

In distributions that are perfectly symmetrical—that is, those in which the left half is a mirror image of the right half—the mean, the median, and (if the distribution is unimodal) the mode all have the same value. It is important to note that the normal distribution (Figure 3.13*f*) falls in this category. Figure 4.3 shows what happens to the mean, median, and mode in **skewed distributions** as compared with the normal distribution. If the mean and median of a distribution have different values, the distribution cannot be symmetrical. Furthermore, the more skewed, or lopsided, the distribution is, the greater the discrepancy between these two measures.

We describe the distribution pictured in Figure 4.3*a* as *skewed to the left or* **negatively skewed**, and that in Figure 4.3*b* as *skewed to the right* or **positively skewed**. In a smooth, negatively skewed distribution (Figure 4.3*a*), the mode has the highest value, and the median falls at a point about two-thirds of the distance between the mode and the mean. The mean, as might be expected, has been specially affected by the fewer but relatively extreme scores in the tail and thus has the lowest value. In a positively skewed distribution (Figure 4.3*b*), exactly the opposite situation is true.[2]

As a consequence, we can use the relative position of the mode, the median, and the mean to determine the direction of skewness without looking at the entire distribution. The magnitude of the discrepancy between the measures gives us a general idea about the degree of skewness as well. For example, doctors at Johns Hopkins School of Medicine accumulated records on 10,000 consecutive births. For women giving birth to their first child, the modal duration of labor was 4 hours, the median 10.6 hours, and the mean 13.0 hours (Guttmacher, 1973).

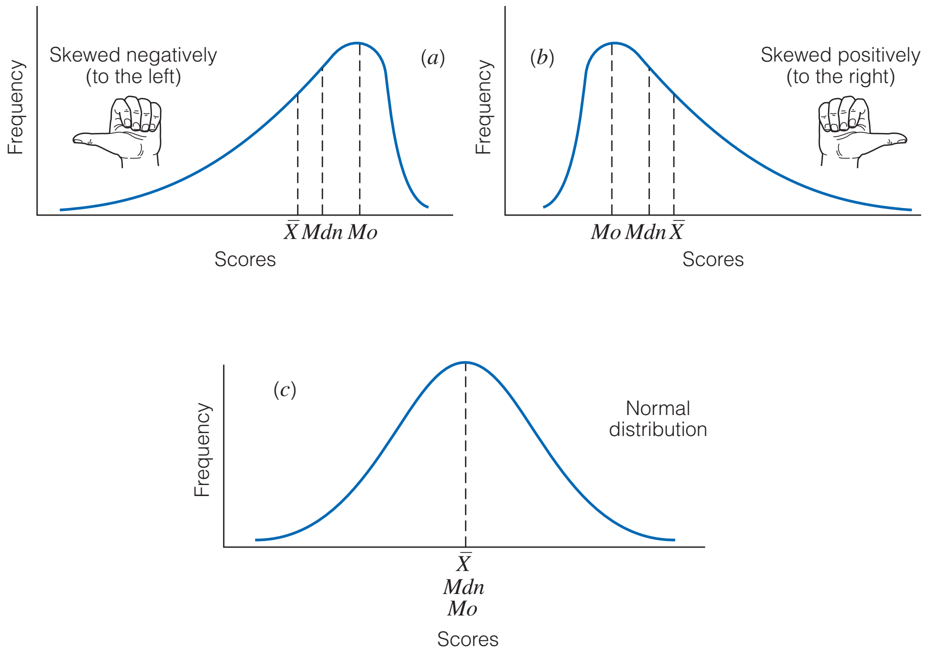


FIGURE 4.3 $\overline{X}$, *Mdn,* and *Mo* in skewed distributions and in the normal distribution.

---

[2] Remember that the position of the score along the horizontal axis indicates the value of these measures of central tendency and not the height of the ordinate erected at these points. The height is simply an indicator of relative frequency of scores at the particular location.

This order matches that for a positive skew, and though we don't know for sure that the distribution had that shape, the supposition makes sense. Probably a few women delivered quickly, most mothers took an intermediate length of time, some continued for many hours, and a few had exceptionally long labors.

For some data we may wish to calculate more than one measure of central tendency. Why? Because each measure conveys a different kind of information. For example, all three measures are useful for summarizing the distribution of American household incomes. In 2005, the income common to the greatest number of households was about $31,000, the median income was $46,326, and the mean income was $63,344. Reporting only one of these measures of central tendency might be misleading and perhaps reflect a bias.

4.8 The Effects of Score Transformations

score transformation a process that changes every score in a distribution to one on a different scale

A **score transformation** is a process that changes every score in a distribution to one on a different scale. In later chapters we will introduce some statistical procedures that require such a change. How will this affect our measures of central tendency?

Consider the following set of scores: 4, 5, 9. Their mean is 6. If we add 10 points to each score, they become 14, 15, 19, and their mean is now 16, 10 points higher. If some constant amount is added to each score in a distribution, the entire distribution is shifted up by the amount of the constant, and the mean will be increased by that same amount. Similarly, if we subtract a constant from each score, the mean will be reduced by that amount. Other measures of central tendency discussed in this chapter are affected in the same way.

linear transformation a score transformation that preserves a straight-line relationship between the original scores and their transformations

Scores may also be transformed by multiplying or dividing each score by a constant. If we multiply the scores 4, 5, and 9 by 2, they become 8, 10, and 18. Whereas the mean of the original set is 6, the mean of the second set is 12 (twice as large). Thus, if we multiply each score by a constant, the mean also multiplies by that amount. Similarly, if we divide each score by a constant, it has the effect of dividing the mean by that amount. If we divide each score in the original set by 10, the resultant scores are .4, .5, and .9, and their mean is .6, a value one-tenth of the original mean. The effect of multiplication or division by a constant holds true also for the mode and median.

The transformations described above all fall under the heading of **linear transformations**, so called because they preserve a proportional, or straight-line, relationship between the original scores and their transformations.

4.9 Summary

A measure of central tendency provides a single summary figure that describes the central location of a distribution of scores. These measures can be particularly useful when we want to compare the performance of two or more groups on the same variable. Three measures are in common use: mode, median, and mean.

The mode is the only measure that can be used for qualitative data. It describes what score or category appears most frequently. The mode sometimes is not a unique point in the distribution, and with grouped data it is more affected by the choice of class interval than are other measures.

The median shows the point along the scale of possible scores that divides the lower half of scores from the upper half. It is another name for P_{50}. The median is responsive to the number of scores above or below it but not to their magnitude and thus is less affected by extreme scores than the mean.

The mean is the sum of all the scores divided by the total number of scores. It is the balance point of the distribution (i.e., the point around which the sum of the negative deviations equals that of the positive deviations). The mean is responsive to the exact position (magnitude) of each score in the distribution and is therefore more sensitive to extreme scores than are the median and the mode.

The three measures of central tendency are affected differently by the symmetry of a distribution. For a bell-shaped, normal distribution, mean, median, and mode have the same value, but for a lopsided (skewed) distribution, their values will differ. Among the three measures, the mean varies the least from sample to sample, which is why it is the measure of choice in most behavioral science applications. Nevertheless, there are some situations (e.g., open-ended distributions and those that are strongly skewed or have a few very deviant scores) in which the median may have significant advantages.

Finally, we note that changing each score by adding or subtracting a constant amount increases or decreases each of the three measures by that same amount. Similarly, if we multiply or divide each score by a constant, it has the effect of multiplying or dividing these measures of central tendency by the same amount.

Mathematical Note

Note 4.1 The Mean as a Balance Point (*Ref.*: Section 4.5)

The algebraic way of saying that the mean is the balance point of the distribution is: $\Sigma(X - \overline{X}) = 0$. The proof is:

$$\begin{aligned}\sum(X - \overline{X}) &= \sum X - n\overline{X}\\ &= \sum X - n\,\frac{\sum X}{n}\\ &= \sum X - \sum X\\ &= 0\end{aligned}$$

N.B.: $\overline{X}$ is a constant when summing over all values in the sample, so $\Sigma\overline{X} = n\overline{X}$

Notice that we made no assumption about the shape of the distribution.

Key Terms, Concepts, and Symbols

measure of central tendency (p. 48)
mode (p. 49)
Mo (p. 49)
median (p. 49)
P_{50} (p. 49)
Mdn (p. 50)
mean (p. 50)
X (p. 51)
Σ (p. 51)
$\overline{X}$ (p. 51)
μ_X (p. 51)
deviation score (p. 52)
$X - \overline{X}$ (p. 52)
outlier (p. 55)
open-ended distribution (p. 55)
skewed distribution (p. 56)
negatively skewed (p. 56)
positively skewed (p. 56)
score transformation (p. 57)
linear transformation (p. 57)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Find the median and the mean for each distribution: (a) 15, 13, 12, 9. (b) 13, 13, 12, 11, 11. (c) 12, 11, 11, 10. (d) 11, 11, 10, 10, 9, 7. (e) 10, 9, 9, 7, 6, 2. (f) 9, 8, 7, 5, 5, 5. (g) 8, 8, 7, 7, 7, 2. (h) 7, 6, 4, 4, 4, 1, 1, 1. (i) 6, 5, 5, 3, 1, 1, 0, 0.

2. Now that you have practiced calculating the mean and median with several small data sets, calculate the mean and median for (a) Data 2A in Chapter 2 (Problem 8), (b) Data 2C (Problem 18), (c) Data 2D (Problem 19), and (d) the data in Table 2.1.

3. Which measure of central tendency is (a) the balance point? (b) best for qualitative data? (c) most sensitive to extreme scores? (d) best for open-ended distributions? (e) sensitive to the value of all scores? (f) responsive to the number of scores below and above it but not to their exact values? (g) lowest in value in negatively skewed distributions? (h) the most resistant to chance sampling variation?

Problems 4 through 8 concern a distribution of 10 scores. Nine of them are 3, 5, 9, 1, 9, 2, 0, 3, and 9. The tenth score, the mystery score, is greater than 5, but it is not 9.

4. On the basis of the information above, is it possible to determine the mode of the distribution of 10 scores? If yes, what is the mode? Explain.

5. On the basis of the information above, is it possible to determine the median of the distribution of 10 scores? If yes, what is the median? Explain.

6. On the basis of the information above, is it possible to determine the mean of the distribution of 10 scores? If yes, what is the mean? Explain.

7. If the mean of the original distribution of 10 scores is 4.8, what is the value of the mystery score? Explain how you calculated it.

8. Assume the mean of the original distribution of 10 scores is 4.8. For each change in the distribution described below, say whether the mean would increase (+), remain unchanged (0), or decrease (−): (a) A value of 6 becomes the 11th score. (b) A value of 4.8 becomes the 11th score. (c) A value of −19 becomes the 11th score. (d) The present value of 2 is removed, leaving nine scores. (e) The mystery value is removed, leaving nine scores. (Remember what was said above about the mystery score.)

9. Look again at the cartoon by Gary Larson. Pat says, "The average American family has 1.5 children." Leslie says, "But there's no such thing as an average person, and there's no such thing as ½ of a child." Mediate the dispute between them.

10. A distribution can be perfectly symmetrical, but have different values for the mode and the median. (a) Explain. (b) Will the mean in this case equal the mode, the median, or something else?

11. At a state university, the first-year class numbers about 360. Until recently, few semifinalists (or finalists) in the National Merit Scholarship program attended. Now the state government has begun to offer its own scholarships to state residents who are National Merit semifinalists—but only if they attend a college in the state. In the first semester of the program, 40 National Merit Scholarship students enter the university. How did the distribution of SAT scores for the first-year class change? Which measure of central tendency changed most? Explain.

12. A researcher finds that the mean of her distribution of measures is 120 and the median is 130. What can you say about the shape of the distribution?

13. A researcher finds that his data form a skewed distribution. He is interested in the bulk of the scores and does not want the few very deviant scores to alter extensively his measure of central tendency. Which measure should he use?

14. A psychologist conducts a learning experiment in which she records the time it takes rats to run through a maze for food reinforcement. A few rats do not complete the maze within 4 minutes, and because of time constraints, she removes them from the maze. Which measure of central tendency would you use for the following distribution of times to traverse the maze?

TIME	*f*
241 sec. and up	4
181–240 sec.	5
121–180 sec.	8
61–120 sec.	5
1–60 sec.	3

15. In the distribution depicted in Figure 4.2, change the "8" to a "2." Recalculate (a) the mode, (b) the median, and (c) the mean. (d) Which measure was most affected by the change? Why?

16. In the distribution depicted in Figure 4.2, omit the "8." Recalculate (a) the mode, (b) the median, and (c) the mean. (d) Which measure was most affected by the change? Why?

17. The mean of a set of scores is 20. What will the mean become if (a) 5 points are added to each score? (b) 15 points are subtracted from each score? (c) Each score is multiplied by 5? (d) Each score is divided by 4? (e) Each score is divided by 2, and 10 is then added to each resulting value? (f) Each score is multiplied by 4, and the resulting values are divided by 5?

18. Some years ago, a newspaper editor claimed that more than half of American families earned a below-average income. Is there any sense in which this claim could be correct? Explain.

19. The National Association of Manufacturers and the steel workers union have both stated the "average" wage for workers in the steel industry. The NAM's figure is higher. Can both be correct? Explain.

CHAPTER 5

Variability and Standard (*z*) Scores

When you have finished studying this chapter, you should be able to:

- Understand what is meant by "measure of variability";
- Calculate the range, the semi-interquartile range, the variance, and the standard deviation;
- Use IBM SPSS to calculate the standard deviation;
- Understand the properties of these four measures of variability and know when it is best to use them;
- Know how each of the measures of variability is affected by adding a constant to each score, by subtracting a constant from each score, or by multiplying or dividing each score by a constant;
- Understand that for tests in the behavioral sciences, there is no standard unit of measurement for the raw scores; and
- Understand the concept of a standard score (z score) and know what effect changing raw scores to z scores has on the mean, the standard deviation, and the shape of the distribution.

Do the scores in a distribution cluster tightly around their central point or do they spread out around it (Figure 5.1*a*)? How much do they spread out? These are questions about the variability among scores. Information about variability is often as important as that about central tendency. We would not be happy with a weight scale that gives our weight correctly on the average, but on any occasion might read 5 pounds lighter or heavier than it should.

measure of variability
a single summary figure that describes the spread of observations within a distribution

Measures of variability *express quantitatively the extent to which the scores in a distribution scatter about or cluster together.* They describe the spread of an entire set of scores; they do not specify how far a particular score diverges from the center of the group. Nor do measures of variability provide information about the shape of a distribution or the central tendency of a group. Figures 5.1*b* and *c*, for example, show that it is possible to have two distributions with equal variability but unequal means (Figure 5.1*b*) or different shapes (Figure 5.1*c*). To describe a distribution adequately, therefore, we usually must provide a measure of central tendency and a measure of variability in addition to describing the shape. We will consider four measures of variability: the range, semi-interquartile range, variance, and standard deviation.

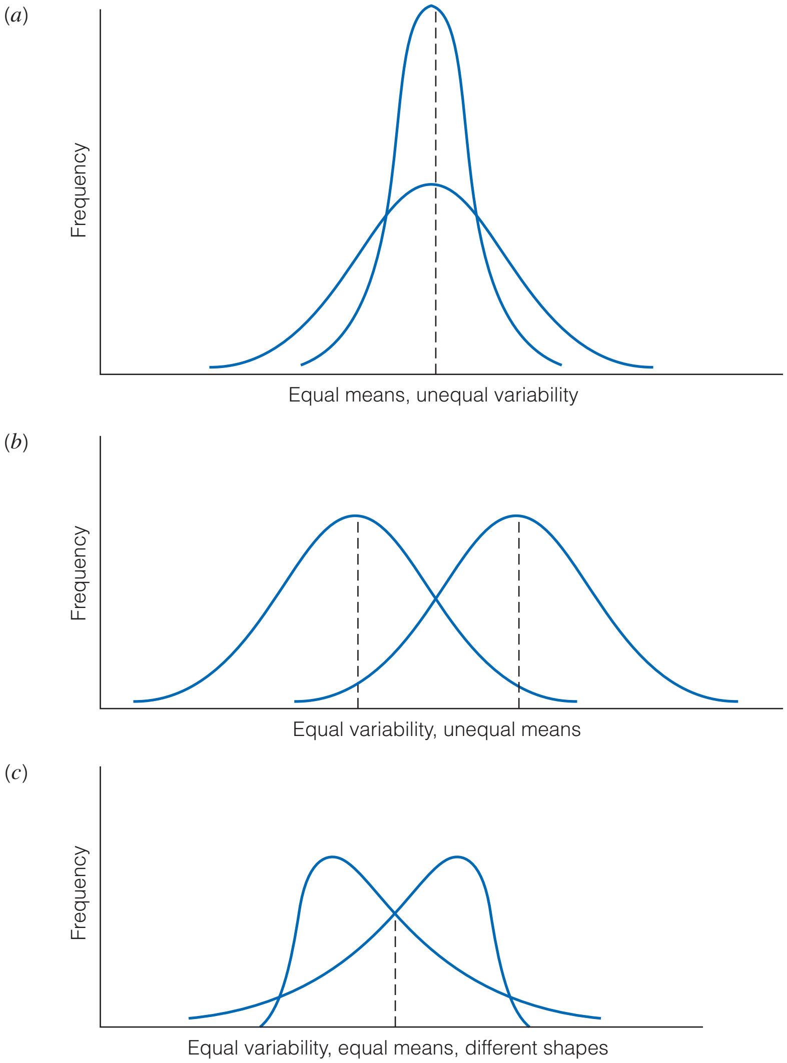


FIGURE 5.1 Differences in central tendency, variability, and shape of frequency distributions (polygons).

5.1 The Range and Semi-Interquartile Range

range
the difference between the lowest score and the highest score in a distribution

The simplest measure of variability is the range, which we met earlier in constructing a frequency distribution. The **range** is the difference between the highest and lowest score in the distribution.[1] Like other measures of variability, the range is a distance and not, like measures of central tendency, a location (on the abscissa). For example, all three of the following simple distributions have the same range (20 score points):

3, 5, 5, 8, 13, 14, 18, 23

37, 42, 48, 53, 57

131, 140, 144, 147, 150, 151

semi-interquartile range
one-half the distance between the first and third quartile points in a distribution

Q
symbol for the semi-interquartile range

quartile points
the three score points that divide a distribution into four parts, each containing an equal number of cases: **Q_1** $(=P_{25})$, **Q_2** $(=P_{50})$, **Q_3** $(=P_{75})$

In a grouped frequency distribution, calculate the range as the difference between the upper apparent limit of the highest interval minus the lower apparent limit of the lowest interval.

Because the range depends on only two scores, the highest and the lowest, it is a crude measure of variability unduly influenced by extreme scores. The **semi-interquartile range**, symbolized by the letter **Q**, depends only on the relatively stable central portion of a distribution—specifically, on the middle 50% of the scores. It is defined as one-half the distance between the first and third quartile points, or

SEMI-INTERQUARTILE RANGE

$$Q = \frac{Q_3 - Q_1}{2} \qquad (5.1a)$$

The **quartile points** are the three score points that divide the distribution into four parts, each containing an equal number of cases. These points, symbolized by Q_1, Q_2, and Q_3, are therefore our old friends, P_{25}, P_{50} (the median), and P_{75}, respectively. We may therefore rephrase Formula 5.1*a* in familiar terms:

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2} \qquad (5.1b)$$

For the data in Figure 5.2, the semi-interquartile range is $(80 - 70)/2 = 5$.

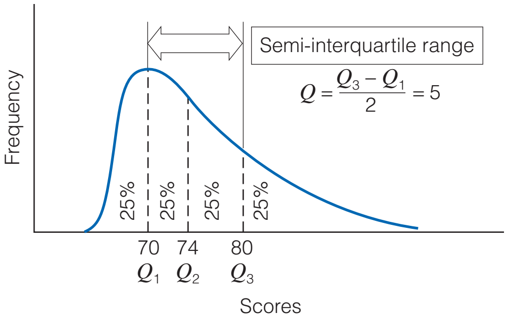


FIGURE 5.2 The semi-interquartile range.

---

[1] Some statistics textbooks calculate the range as the difference between the upper real limit of the highest score and the lower real limit of the lowest score. For simplicity's sake, we will use raw scores.

5.2 Deviation Scores

$(X - \overline{X})$ a deviation score; shows by how much a score differs from the mean

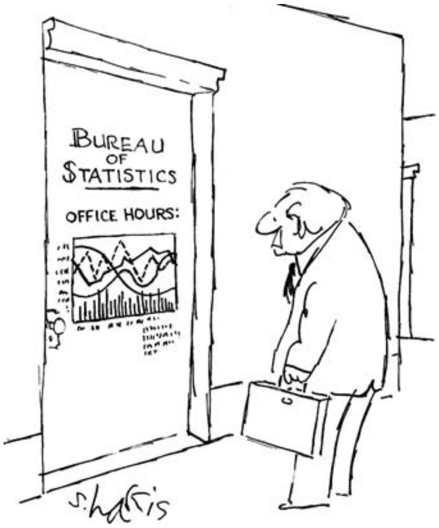
You were introduced to deviation scores in Section 4.5. Recall that a **deviation score** expresses the location of a score by indicating how many score points it lies above or below the mean of the distribution. In symbols, the deviation score is defined as **$(X - \overline{X})$**. Consider the following raw scores (X) and their corresponding deviation scores ($X - \overline{X}$):

X	$(X - \overline{X})$	
1	$(1 - 4) = -3$	
5	$(5 - 4) = +1$	
7	$(7 - 4) = +3$	$\overline{X} = \dfrac{16}{4} = 4$
3	$(3 - 4) = -1$	
$\sum X = 16$		

When we subtract the mean ($\overline{X} = 4$) from each of the raw scores, the resulting deviation scores state the position of the scores *relative to the mean*. A plus sign (+) indicates that the original raw score is greater than the mean, whereas a minus sign (−) indicates that it is less than the mean. For example, the raw score of 7 becomes $(7 - 4) = +3$, which says that this score is 3 points above the mean of 4. Similarly, the raw score of 3 becomes $(3 - 4) = -1$, indicating that its location is 1 point below the mean. We now turn to two important measures of variability based on the concept of the deviation score.

5.3 Deviational Measures: The Variance

Because deviation scores indicate the distances of the raw scores from the mean (the most frequently used measure of central tendency), it might appear that an attractive measure of variability could be obtained by calculating the mean of the deviation scores. This would require finding their sum. Unfortunately, as we learned in Section 4.5, the sum of the deviations from the mean *always* equals zero. Look again at the four deviation scores in the previous section ($-3, +1, +3, -1$). When we add them together, $\Sigma(X - \overline{X}) = 0$.



Source: © Sidney Harris, ScienceCartoonsPlus.com

We could drop the signs of the deviation scores before taking their mean, but this yields a measure with such awkward mathematical properties that it is useless in inferential statistics. Of course, another way to get rid of the signs is to square the deviation scores. This is what is done in a very useful measure of variability called the variance.

variance the mean of the squares of the deviation scores

σ^2 symbol for the variance of a population

The **variance** is defined as the mean of the squares of the deviation scores (or, as many prefer, the "average squared deviation from the mean"). The symbol for the variance of a population is **σ^2** (the lowercase Greek letter sigma) and that for a sample is **S^2**. The defining formulas for the variance are as follows:

$\sigma_X^2 = \dfrac{\Sigma(X - \mu_X)^2}{N}$ definitional formula for the variance of a population (X)

DEFINITIONAL FORMULA FOR
VARIANCE OF A POPULATION (X)

$$\sigma_X^2 = \frac{\sum(X - \mu_X)^2}{N} \quad \text{or} \quad \frac{SS_X}{N} \qquad (5.2a)$$

DEFINITIONAL FORMULA FOR
VARIANCE OF A SAMPLE (X)

$$S_X^2 = \frac{\sum(X - \overline{X})^2}{n} \quad \text{or} \quad \frac{SS_X}{n} \qquad (5.2b)$$

S^2 symbol for the variance of a sample

$S_X^2 = \dfrac{\Sigma(X - \overline{X})^2}{n}$ definitional formula for the variance of a sample (X)

The numerator of the variance, $\Sigma(X - \overline{X})^2$, is used so frequently in other statistical formulas that it has its own abbreviated name: the **sum of squares**, which is short for the *sum of the squared deviations from the mean*. It is often symbolized by ***SS***. The steps in calculating the variance by Formula 5.2*b* are summarized as follows and are illustrated in Table 5.1.

sum of squares short for "sum of the squared deviations from the mean"

Step 1: Record each score.

Step 2: Calculate $\overline{X}$.

Step 3: Obtain deviation scores by subtracting $\overline{X}$ from each value of X (as a check on your work, remember that $\Sigma(X - \overline{X})$ should equal zero).

SS symbol for sum of squares

$SS_X = \Sigma(X - \overline{X})^2$ formula for the sum of squares of a sample (X)

TABLE 5.1 *Calculation of the Variance: Deviation-Score Method*

① X	③ $(X - \overline{X})$	④ $(X - \overline{X})^2$
32	$32 - 50.6 = -18.6$	345.96
71	$71 - 50.6 = +20.4$	416.16
64	$64 - 50.6 = +13.4$	179.56
50	$50 - 50.6 = -0.6$	.36
48	$48 - 50.6 = -2.6$	6.76
63	$63 - 50.6 = +12.4$	153.76
38	$38 - 50.6 = -12.6$	158.76
41	$41 - 50.6 = -9.6$	92.16
47	$47 - 50.6 = -3.6$	12.96
52	$52 - 50.6 = +1.4$	1.96
$\sum X = 506$	$\sum(X - \overline{X}) = 0$	⑤ $\sum(X - \overline{X})^2 = 1368.40$

② $\overline{X} = \dfrac{506}{10} = 50.6$

⑥ $S_X^2 = \dfrac{\sum(X - \overline{X})^2}{n} = \dfrac{1368.40}{10} = 136.84$

Step 4: Square each deviation score.

Step 5: Sum the values of the squared deviation scores to get the sum of squares (*SS*).

Step 6: Divide the sum of squares by *n* as in Formula 5.2*b*.

5.4 Deviational Measures: The Standard Deviation

The variance is frequently used in advanced statistical procedures and especially in inferential statistics. However, for basic description, it has a fatal flaw: Its calculated value is expressed in *squared* units of measurement. (If the scores are weights in pounds, the variance will be a certain quantity of squared pounds.) Consequently, it is of little use in descriptive statistics. However, this is easily remedied. By taking the square root of the variance, we return to the original units of measurement and thereby obtain a widely used index of variability called the **standard deviation**. Although it is *not technically correct*, it is easier for beginning students to think of the standard deviation as "the average amount that scores deviate from the mean." When there is wider scatter among the raw scores, there are bigger deviations from the mean, and thus the "standard" (typical, representative) deviation increases.

standard deviation
the square root of the variance

σ
symbol for the standard deviation of a population

The defining formulas for the standard deviation are:

DEFINITIONAL FORMULA FOR
STANDARD DEVIATION OF A POPULATION

$$\sigma_X = \sqrt{\frac{\sum(X-\mu_X)^2}{N}} \quad \text{or} \quad \sqrt{\frac{SS_X}{N}} \qquad (5.3a)$$

DEFINITIONAL FORMULA FOR
STANDARD DEVIATION OF A SAMPLE

$$S_X = \sqrt{\frac{\sum(X-\overline{X})^2}{n}} \quad \text{or} \quad \sqrt{\frac{SS_X}{n}} \qquad (5.3b)$$

$$\sigma_X = \sqrt{\frac{\Sigma(X-\mu_X)^2}{N}}$$

definitional formula for the standard deviation of a population (*X*)

S
symbol for the standard deviation of a sample

As you can see, the symbols for the standard deviation of a population (**σ**) and a sample (**S**) are the same as those used for the variance, but without the square root sign. To obtain the standard deviation of the sample of scores in Table 5.1, we simply take the square root of the variance:

$$S_X = \sqrt{\frac{\sum(X-\overline{X})^2}{n}} = \sqrt{\frac{1368.40}{10}} = \sqrt{136.84} = 11.7$$

$$S_X = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$$

definitional formula for the standard deviation of a sample (*X*)

5.5 Calculation of the Variance and Standard Deviation: Raw-Score Method

When we calculate the variance or standard deviation by the deviation-score formula, calculating *SS* directly from deviation scores is very time consuming. Minium's Second Law of Statistics states: "*For computation, if there is a choice between a deviation-score method and some other* *method, choose the other method."* Without the bother of changing the raw scores to deviation scores, we can find SS by the equation:[2]

$\boldsymbol{SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{n}}$
computational formula for the sum of squares

COMPUTATIONAL
(RAW-SCORE) EQUIVALENT OF SS

$$SS_X = \sum(X - \overline{X})^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n} \qquad (5.4)$$

$\boldsymbol{S_X^2 = \frac{SSx}{n}}$
another way of expressing the formula for the variance of a sample (X)

Remember, $S_X^2 = SS_X/n$ (Formula 5.2*b*), and $S_X = \sqrt{SS_X/n}$ (Formula 5.3*b*). Derivation of Formula 5.4 appears in Note 5.1 at the end of the chapter. The steps in calculating the standard deviation by the raw-score method are summarized as follows and are illustrated for a set of 10 scores in Table 5.2:

$\boldsymbol{S_X = \sqrt{\frac{SSx}{n}}}$
another way of expressing the formula for the standard deviation of a sample (X)

Step 1: Record each score.
Step 2: Record the square of each score.
Step 3: Calculate ΣX and ΣX^2.

TABLE 5.2 *Calculation of the Standard Deviation: Raw-Score Method*

① X	② X^2
32	1,024
71	5,041
64	4,096
50	2,500
48	2,304
63	3,969
38	1,444
41	1,681
47	2,209
52	2,704
③ $\sum X = 506$	③ $\sum X^2 = 26{,}972$

Calculation of SS_X:

④ $$SS_X = \sum(X - \overline{X})^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$
$$= 26{,}972 - \frac{(506)^2}{10}$$
$$= 26{,}972 - 25{,}603.6$$
$$= 1368.4$$

Calculation of S_X:

$$S_X^2 = \frac{SS_X}{n}$$
$$S_X = \sqrt{\frac{SS_X}{n}}$$
⑤ $$= \sqrt{\frac{1368.4}{10}}$$
$$= \sqrt{136.84}$$
⑥ $$= 11.7$$

---

[2] For the remainder of the chapter, we will present formulas appropriate for samples in most examples; the modifications necessary for a population should be readily apparent.

Step 4: Calculate SS_X in accord with Formula 5.4. In doing so, distinguish carefully between ΣX^2 and $(\Sigma X)^2$. To find the former, we square each score and then sum the squares. To find the latter, we sum the scores and then square the sum. The results are not the same. Keep in mind that the sum of squares *must* be a positive number no matter what the raw score values are.

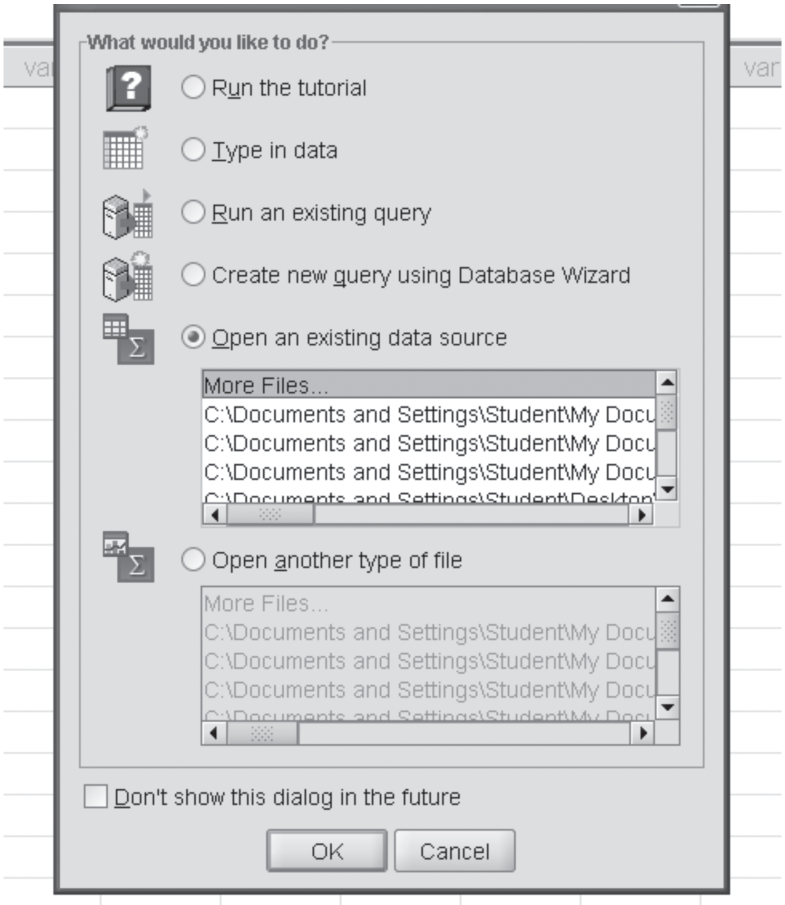
Step 5: Substitute the numerical values for SS and n in the formula for S.

Step 6: Complete the calculation of S.

5.6 Calculation of the Standard Deviation with IBM SPSS (formerly SPSS)

Once IBM SPSS is installed, you may open it by double-clicking an icon that appears on the desktop, or by clicking on the Start button and selecting the IBM SPSS program.

When opening IBM SPSS, a dialog box will appear with the following question: "What would you like to do?" (see Figure 5.3). There are various options from which to choose, including "Run the tutorial," "Type in data," and "Open an existing data source." Because data will be entered for the first time, you should select "Type in data" and then click the OK button.



What would you like to do?

○ Run the tutorial

○ Type in data

○ Run an existing query

○ Create new query using Database Wizard

● Open an existing data source

More Files...
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\Desktop

○ Open another type of file

More Files...
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\My Docu
C:\Documents and Settings\Student\My Docu

☐ Don't show this dialog in the future

FIGURE 5.3 Dialog box when first opening IBM SPSS.

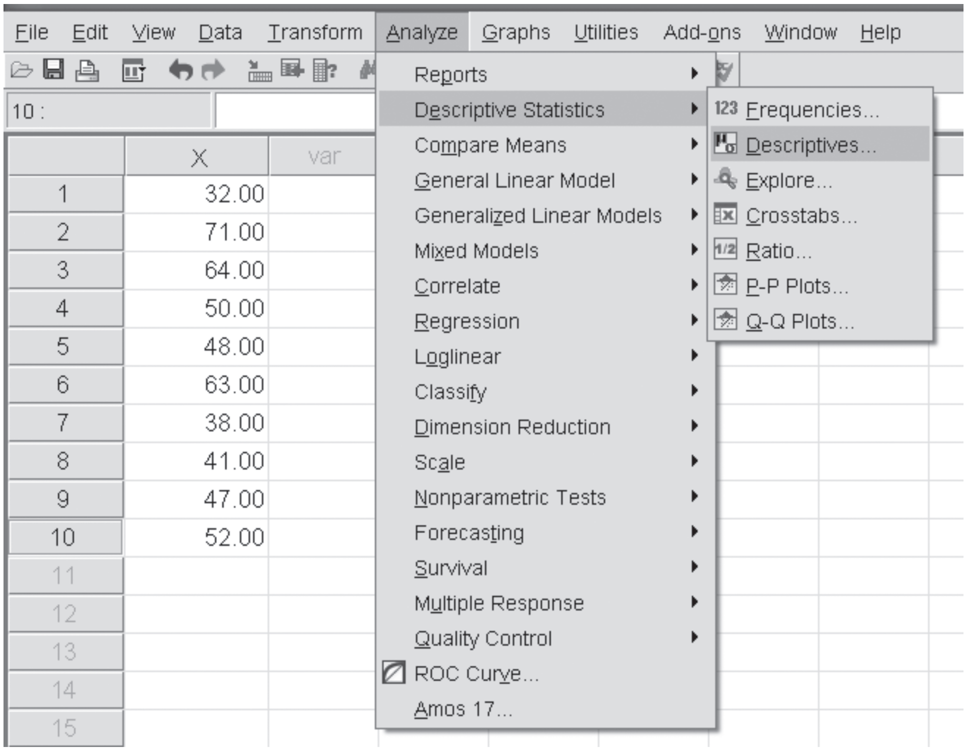
Notice that the menu at the top of the interface includes **Menu**, **Edit**, and **View**. This may appear very familiar to you because many applications include such menu labels. Note also that at the bottom left of the interface there are two tabs, one labeled **Data Editor** and the other labeled **Variable View**. By default, IBM SPSS starts off in the **Data Editor**. However, to switch between the **Data Editor** and **Variable View**, you can click on the appropriate tab.

Using the same scores that are shown in Tables 5.1 and 5.2, enter your data in the **Data Editor**. The rows are numbered and are often referred to as cases. The columns are used for variables and these are not labeled. However, you can enter variable names by clicking on the **Variable View** tab. Now click on the **Variable View** tab and type an 'X' in the first row, under the column labeled Name.

After you have entered the data in the **Data Editor**, you can obtain the mean and standard deviation by clicking **Analyze** and then selecting **Descriptive Statistics** followed by **Descriptives**. The screenshot of the data as it would appear in the Data **Editor** of IBM SPSS will look like Figure 5.4. Notice the menus that would appear after clicking **Analyze** → **Descriptive Statistics** → **Descriptives**.

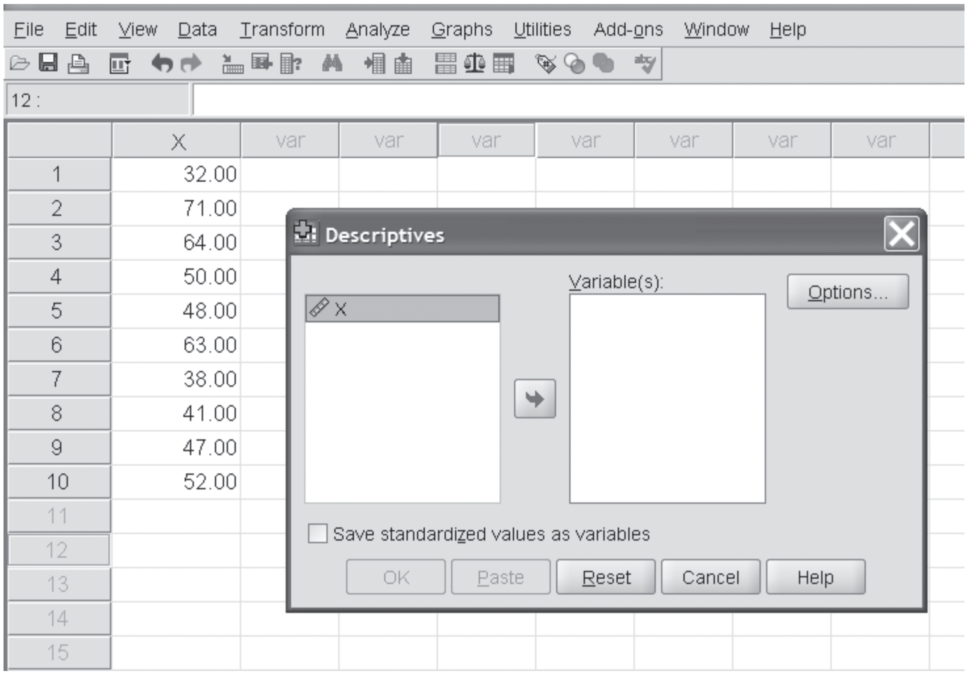
Next, a dialog box will appear as in Figure 5.5. All variables in your data file will appear in the box on the left. Now highlight the variable to be analyzed and move it to the box on the right by clicking on the arrow button that appears in the center of the dialog box. Figure 5.6 is a screenshot after the X variable has been moved to the box on the right. Click the OK button.

As part of the output, IBM SPSS will produce the relevant syntax used to conduct an analysis, followed by the results of the statistical analysis. The results typically appear in tabular form. You can see the IBM SPSS output in Table 5.3.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
10 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
P-P Plots...
Q-Q Plots...

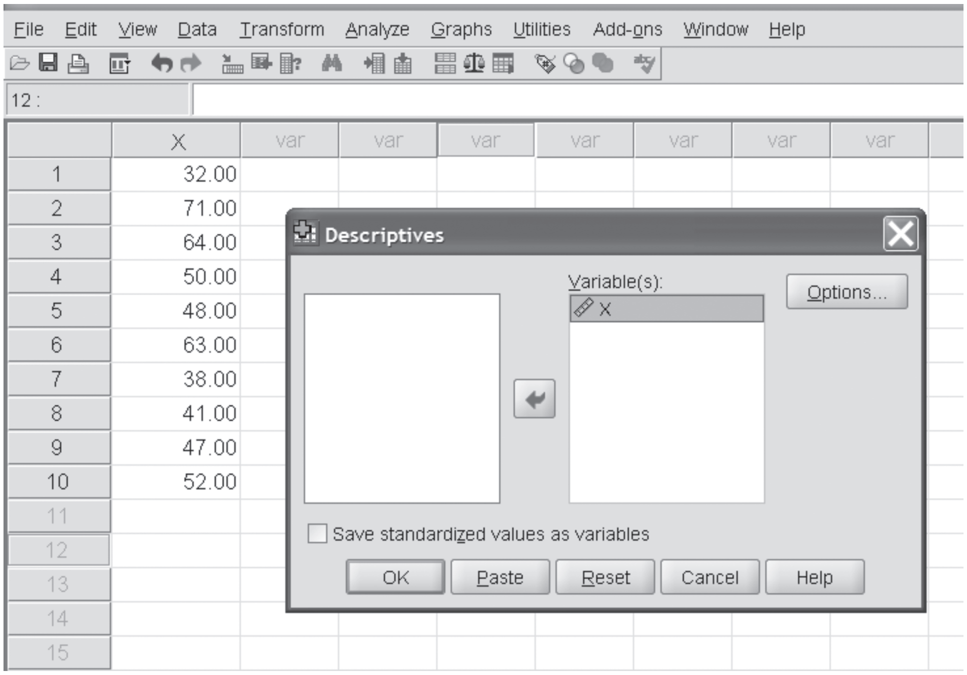
	X	var
1	32.00	
2	71.00	
3	64.00	
4	50.00	
5	48.00	
6	63.00	
7	38.00	
8	41.00	
9	47.00	
10	52.00	
11		
12		
13		
14		
15		

FIGURE 5.4 IBM SPSS menus when calculating descriptive statistics.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
12 :
var
Descriptives
Variable(s):
Options...
X
☐ Save standardized values as variables
OK
Paste
Reset
Cancel
Help

	X
1	32.00
2	71.00
3	64.00
4	50.00
5	48.00
6	63.00
7	38.00
8	41.00
9	47.00
10	52.00
11	
12	
13	
14	
15	

FIGURE 5.5 IBM SPSS dialog box for descriptive statistics.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
12 :
var
Descriptives
Variable(s):
X
Options...
☐ Save standardized values as variables
OK
Paste
Reset
Cancel
Help

	X
1	32.00
2	71.00
3	64.00
4	50.00
5	48.00
6	63.00
7	38.00
8	41.00
9	47.00
10	52.00
11	
12	
13	
14	
15	

FIGURE 5.6 IBM SPSS dialog box for descriptive statistics after selecting a variable.

You will notice that the output in Table 5.3 shows that the standard deviation is 12.33063 while Table 5.2 shows it to be 11.7. This is because IBM SPSS divides the sum of squares (*SS*) by $n - 1$ rather than n. See the Point of Controversy regarding division by n versus $n - 1$. To get 11.7, square 12.33063 and multiple by 9/10. Then take the square root.

TABLE 5.3 *IBM SPSS Output of Descriptive Statistics*

```
DESCRIPTIVES VARIABLES = X
/STATISTICS = MEAN STDDEV MIN MAX.
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X	10	32.00	71.00	50.6000	12.33063
Valid N (listwise)	10				

Point of Controversy

Calculating the Sample Variance: Should We Divide by *n* or (*n* − 1)?

In Section 5.3, we defined the variance as the mean of the squares of the deviation scores. To calculate the variance of a *population* we divided the sum of the squared deviations from the mean (*SS*) by *N*, the number of scores in the population. To calculate the variance of a *sample* we (like many statisticians) also use this formula (substituting *n* for *N*). However, some prefer to define the variance of a sample as $\Sigma(X - \overline{X})^2/(n - 1)$. If your calculator has a program for the sample variance, it may also use this latter formula. The use of two different formulas for the sample variance is, of course, confusing to students who have compared textbooks that differ in this respect.

Why do some statisticians divide the sample sum of squares by *n* and others by (*n* − 1)? The answer has to do with the reason an investigator obtains a sample for study in the first place. Rarely are we interested in the sample itself. Instead, we obtain samples and calculate statistics to make inferences about population parameters. Due to chance factors in drawing a sample, we do not expect our sample statistic (e.g., sample mean, sample variance) to exactly equal the population parameter (e.g., population mean, population variance), but we hope it is close. As we will learn later, the sample mean is an example of what we call an "unbiased" statistic. Although the mean of a particular sample is probably not going to equal the mean of the population from which the sample was drawn, there is no bias in the direction of the error. There is an equal likelihood that it will fall below or above the population mean.

The sample variance, however, is not an unbiased statistic. *It tends to be an underestimate of the population variance*. The reason is that the sum of the squared deviations from the mean of any set of scores is less than the sum of squares calculated from any other point. (We discuss this in Section 5.8 and prove it in Note 5.2.) Because the sample mean is rarely exactly equal to the population mean, the sum of squares calculated from a sample mean is usually less than if it had been calculated from the population mean (which is one of those "any other points"). To correct for this underestimation in the numerator of the formula for the sample variance, statisticians divide *SS* by (*n* − 1) instead of *n*. The result is no longer the variance of the *sample*, but what is called the *unbiased estimate of the population variance*. The value of $\Sigma(X - \overline{X})^2/(n - 1)$ will still not be equal to the population variance in most cases, but instead of a biased estimate (tending to be an underestimate), there is an equal likelihood of the value falling below or above the value of the population variance.

So, which formula is correct? Actually, they both are. To obtain the variance for any given set of scores, we always divide by *n*. Thus, the variance of a sample is calculated by dividing by *n*. Ultimately, however, we are interested in estimating the population variance, and to get an unbiased estimate of that, we divide by (*n* − 1) (Cormack & Mantel, 1990). We will discuss this further when we get to inferential statistics.

5.7 Properties of the Range and Semi-Interquartile Range

The range is easier to compute than the other measures of variability and its meaning is direct. Therefore, *the range is ideal for preliminary work or in other circumstances where precision is not an important requirement*. The range, however, has a major shortcoming as a measure of variability. Only the two outermost scores of a distribution affect its value; the remainder could lie anywhere between them. In fact, a single outlier (an aberrant score) has a more substantial effect on the range than on the other measures.[3] Thus, *the range is not sensitive to the total condition of the distribution*.

The semi-interquartile range is closely related to the median because both are defined in terms of percentile points of the distribution. The median is responsive to the *number* of scores lying below it rather than to their exact position, and P_{25} and P_{75} are points defined in a similar way. We may therefore expect the median and the semi-interquartile range to have properties in common. For example, Q_3 is sensitive only to the number of scores that lie above it. Consequently, all scores above Q_3 are of equal importance in the determination of the semi-interquartile range, and a very extreme score counts no more and no less than a moderately extreme one in its determination. Therefore, *the semi-interquartile range is less sensitive to the presence of a few very extreme scores than is the standard deviation*. If the distribution is open-ended (see Section 4.6), it is not possible to calculate the standard deviation or the range without making additional (and probably hazardous) assumptions. *With open-ended distributions, the semi-interquartile range may be the only measure that is reasonable to compute.* Unless more than a quarter of the scores lie in the indeterminate end category, we may compute the semi-interquartile range as we normally would.

5.8 Properties of the Standard Deviation

The range and the semi-interquartile range are intuitively easy to understand as measures of variability. It is ironic, therefore, that the standard deviation is more important in statistical work than the others, but less easy to understand. In many ways, the properties of the standard deviation are related to those of the mean. *The standard deviation, like the mean, is responsive to the exact position of every score in the distribution.* Because it is calculated by taking deviations from the mean, if a score is shifted to a position more deviant from the mean, the standard deviation will increase. If the shift is to a position closer to the mean, the standard deviation decreases.

The standard deviation is more sensitive than the semi-interquartile range to the presence or absence of scores that lie at the extremes of the distribution. Because of this sensitivity, *the standard deviation may not be the best choice among measures of variability when the distribution contains a few very extreme scores or when the distribution is badly skewed.*

When we calculate deviations from the mean, the sum of squares of these values is *smaller* than if they had been taken about any other point. Putting it another way,

$$\sum (X - \overline{X})^2 < \sum (X - \text{any other score})^2$$

To illustrate this property, consider the scores 2, 3, 5, and 10, which have a mean of 5. Calculate the sum of squares of the deviation scores, taking the deviations from the mean. Then recalculate, this time taking the deviations from 3. Repeat again, taking the deviations from 2. Do it again from 10. What do you find?

---

[3] We introduced the term *outlier* in Chapter 4 and to this point have referred to it as an extreme or aberrant score. Some statisticians state that a score must be at least 2.5 standard deviations from the mean to be called an outlier.

We provide proof of this proposition in Note 5.2 at the end of the chapter. This property may seem to be an oddity that only a mathematician could love. In fact, it is an extremely important principle, one that reappears frequently in the further study of statistics (see, for example, Section 8.2). For now, note that this gives us another way to define the mean: *It is the point about which the sum of squares of the deviation scores is a minimum.*

When you begin to study additional methods of statistical analysis, you will better understand the importance of the standard deviation. *The standard deviation appears explicitly or lies embedded in many procedures of both descriptive statistics and inferential statistics.*

5.9 How Big Is a Standard Deviation?

What does it mean to say that a score is 1 standard deviation away from the mean? Two standard deviations? Distributions differ in their shape, so we cannot make generalizations that will apply in all cases. However, as an example, we will examine the normal distribution. We define a normal distribution in Chapter 6—all you need to know for now is that normal curves are bell-shaped (as shown in Figure 3.13*f*). In any normal distribution, the interval:

$\mu \pm 1\sigma$	contains about 68% of the scores
$\mu \pm 2\sigma$	contains about 95% of the scores
$\mu \pm 3\sigma$	contains about 99.7% of the scores

We show these relationships in Figure 5.7.

By now you should be starting to appreciate what we mean when we say that a score is 1 or 2 standard deviations from the mean. Although we usually do not think of 2 as a large number, in terms of standard deviations from the mean, it is a rather large distance.

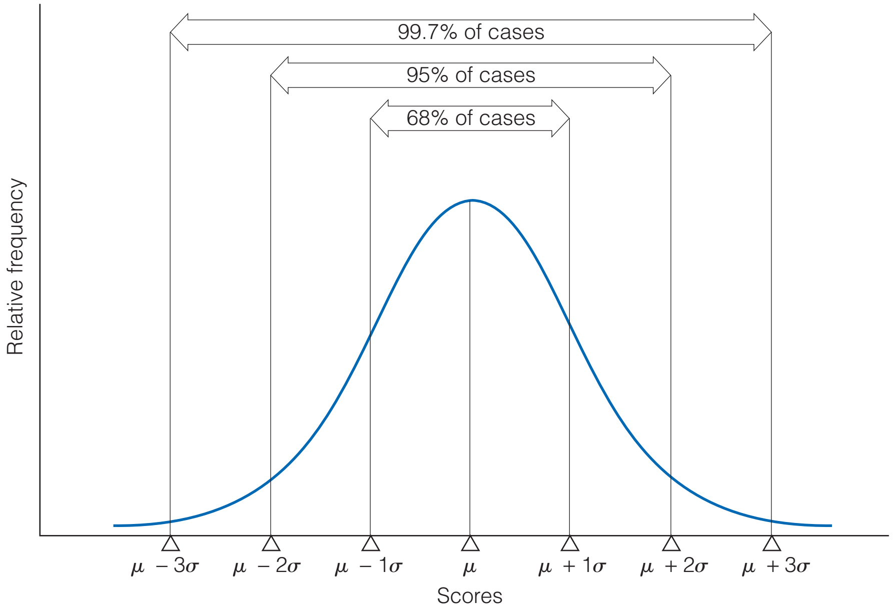


FIGURE 5.7 Relative frequencies of cases contained within certain limits in the normal distribution.

5.10 Score Transformations and Measures of Variability

In Section 4.8, we found that certain score transformations had simple and predictable effects on measures of central tendency. What are the effects on measures of variability?

Consider the following set of scores: 12, 13, and 15. The distance between the first and second scores is 1 point, and that between the second and third scores, 2 points. If we add 10 points to each score, they become 22, 23, and 25. The interscore distances, however, remain the same. If we subtract 5 points from each score, they become 7, 8, and 10. Again, the interscore distances remain the same. *If we add a constant to each score in the distribution or subtract a constant from each score, it does not affect any of the measures of variability described in this chapter.*

Scores may also be transformed by multiplying each score by a constant or by dividing each score by a constant. If, for example, we multiply the scores 12, 13, and 15 by 2, they become 24, 26, and 30. In their original form, the interscore distances are 1 and 2 points, respectively. In their altered form, they are doubled, becoming 2 and 4 points. If we multiply the original scores by 10, the scores become 120, 130, and 150, and the interscore distances are now 10 points and 20 points. If we alter the original set of scores by dividing each score by 10, the scores become 1.2, 1.3, and 1.5. The interscore distances are now .1 and .2, one-tenth of their original values. *When scores are multiplied or divided by a constant, the resultant measure of variability is also multiplied or divided by that same constant.* This applies to all measures of variability discussed in this chapter except the variance.[4]

5.11 Standard Scores (*z* Scores)

Your best friend has just received his grade on a history examination. "How did you do?" you ask. "I got a score of 123," he replies. You do not know whether to be excited or to sympathize. "What was the class mean?" you inquire. "One hundred and five," he responds. You are happy for him, but want to know more: "What was the range?" "The low score was 80, and the high score was 125," he replies.

This example illustrates *the lack of unequivocal meaning that is typical for measurement in the behavioral sciences.* A raw score, by itself, tells us nothing. We need a frame of reference to decide whether a score indicates a good performance or a poor one. In the example, learning the value of the mean of the distribution helped you to understand your friend's level of performance, and information about the range enabled you to see that his performance was very nearly the best in the class. You were able to interpret the raw score by relating it to a measure of central tendency and a measure of variability.

The solution to the problem of making scores meaningful involves providing an adequate frame of reference. If we measure the length of an object in feet, we immediately understand the measurement because the length obtained is the same whether measured by your yardstick or mine and also because we have some understanding of how long a foot is. In mental measurement, the situation is different. In measuring achievement in history, or intelligence, or mechanical aptitude, there is no standard unit of measurement for the raw scores. The size and spread of the scores depend on the number of test items, the number of points assigned to each item, and the difficulty of the items. There are two additional problems. First, a test of intelligence, for example, is only a *sample* of questions designed to tap this characteristic. Another test of the same characteristic will therefore constitute a somewhat different sample and will have a somewhat different outcome. Second, because we are not in perfect agreement as to what intelligence is, two test constructors may sample different aspects of intelligence. All of these factors indicate that, unlike the measurement of distance, one "yardstick" of intellectual functioning is not necessarily the same as another.

---

[4] Variance is expressed in terms of squared units of measurement. Consequently, if each score is multiplied or divided by a constant C, the variance is multiplied or divided by C^2.

How, then, can we answer the question? One possibility is to use percentile ranks (Section 2.9). Another, often more useful approach is to convert the raw scores to standard scores. A standard score, or ***z* score**, states the position of a score in relation to the mean of the distribution, using the standard deviation as the unit of measurement. For example, in a distribution that has $\mu = 100$ and $\sigma = 20$, a score of 120 may be expressed as a z score of $+1$, indicating that the score is 1 standard deviation above the mean of the distribution. Similarly, a score of 60 in the same distribution is expressed as a z score of -2.0, because it is 2 standard deviations below the mean. The formula for a z score is:

***z* score**
states how far away a score is from the mean in standard deviation units; one type of standard score

DEFINITIONAL FORMULA FOR
z SCORE IN A POPULATION

$$z = \frac{X - \mu_X}{\sigma_X} \tag{5.5a}$$

DEFINITIONAL FORMULA FOR
z SCORE IN A SAMPLE

$$z = \frac{X - \overline{X}}{S_X} \tag{5.5b}$$

$$z = \frac{X - \mu_X}{\sigma_X}$$

formula for a z score in a population

$$z = \frac{X - \overline{X}}{S_X}$$

formula for a z score in a sample

The z score makes it possible, under *some* circumstances, to compare scores that originally had different units of measurement because after conversion each has a common unit of measurement. For example, how do we compare the performance on a history and a chemistry exam? That is like trying to compare apples to oranges.

Suppose that Helen, a fifth grader, earned an 80 on her American history exam and a 65 on her arithmetic exam in the *same* class. Is she doing better in history or arithmetic? First, we need to know how the rest of Helen's class did on the two exams. Suppose that the mean score on the history exam was 65 and that the mean score on the arithmetic exam was 50. Helen thus scored 15 points higher than the mean for the class on both exams. Can we conclude that she performed equally well in both subjects? We still do not know because we don't know whether 15 points on the history exam is the same as 15 points on the arithmetic exam. Suppose also that the standard deviation on Helen's history exam was 15, but that the standard deviation on her arithmetic exam was only 7.5. Now we can determine in which subject she did better. Helen's history score of 80 has a z value of $+1.0$, because it was $+1.0$ standard deviations greater than the mean for her class $[(80 - 65)/15]$. Her arithmetic score has a z value of $+2.0$, because it was $+2.0$ standard deviations greater than the mean $[(65 - 50)/7.5]$. In relation to the rest of her class, Helen did better in arithmetic.

In this example, the history and arithmetic scores were earned by the same students. It may not be appropriate to compare scores from *different* groups. To do so, it is necessary that the two groups be comparable (i.e., that the two distributions of scores have roughly the same shape).

A z score is of great importance in statistical work, and it will appear again and again in this book. You will find it a great help to fully understand the following verbal definition of a z score: *the* z *score states by how many standard deviations the corresponding raw score lies above or below the mean of the distribution.*

There are three important properties of z scores (see if you can derive them from the definition):

1. ***The mean of any distribution of scores converted to z scores is always* 0: $\mu_z = 0$.** Because the mean of the raw scores is the score that is no standard deviations away from the mean, it is represented by $z = 0$.
2. ***The standard deviation of any distribution expressed in z scores is always* 1: $\sigma_z = 1$.** In calculating z scores, the standard deviation of the raw scores is the unit of measurement, so the standard deviation of the z scores is unity.
3. ***Transforming raw scores to z scores changes the mean to* 0 *and the standard deviation to* 1*, but it does not change the shape of the distribution.*** (The reason is that for each raw score, we first subtract a constant—the mean—and then divide by another constant—the standard deviation.) By "shape of the distribution," we really mean the proportional relation that exists among the distances between the scores. For example, consider the following scores and their z score equivalents:

   Four scores from a distribution where $\mu = 80$ and $\sigma = 10$:

   70 75 85 100

   Interscore differences among raw scores: 5 10 15

   Equivalent z scores: -1 $-\frac{1}{2}$ $+\frac{1}{2}$ $+2$

   Interscore differences among z scores: $\frac{1}{2}$ 1 $1\frac{1}{2}$

Notice the relation between the interscore distances exhibited for the raw scores and the z scores. The proportional magnitude of difference between successive raw scores, 5:10:15, is exactly the same in z score terms, $\frac{1}{2}$:1:$1\frac{1}{2}$. Thus, if a distribution of raw scores is skewed right, with the top score twice as far from the mean as is the bottom score, the z scores will still be skewed right, and the absolute value of the top z score will be twice as large as that of the bottom z score. Figure 5.8 illustrates what happens when a set of raw scores having a

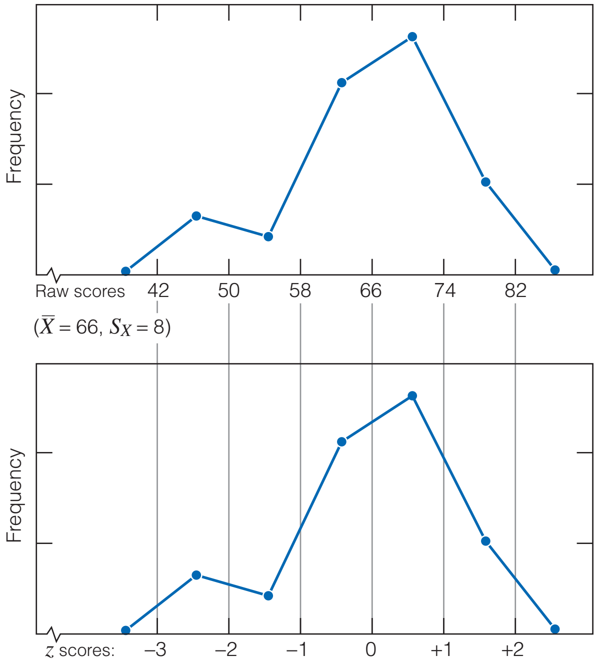


FIGURE 5.8 Comparison of the shapes of a raw-score distribution and the same distribution converted to z scores.

TABLE 5.4 *Some Common Standard Scores and Their Uses*

TYPE	MEAN	STD DEV	EXAMPLE OF USE
1	50	10	Minnesota Multiphasic Personality Inventory (MMPI) California Psychological Inventory (CPI)
2	100	16	Stanford–Binet Intelligence Scale
3	100	15	Wechsler Intelligence Scale IQs (WISC-R)

mean of 66 and a standard deviation of 8 is expressed as a set of z scores (with a mean of zero and a standard deviation of 1).

standard scores all have a fixed mean and fixed standard deviation; a z score is one type

The z score is only one type of **standard score**. Other varieties have been devised. *Each of them is like the* z *score in that it has a fixed mean and standard deviation;* this is what is "standard" about them. Table 5.4 gives three common examples. The fundamental characteristic of all types of standard scores is that they, like the z score, locate the raw score by stating by how many standard deviations it lies above or below the mean of the distribution. For example, on the scale having a mean of 50 and a standard deviation of 10, a score of 40 is 10 points below the mean, and because the value of the standard deviation is 10 points, we know that this score is 1 standard deviation below the mean. On the remaining scales in Table 5.4, a score of 84 and 85, respectively, each falls 1 standard deviation below the mean. We wanted to introduce you to the fact that there are several types of standard scores. However, in this book we will focus exclusively on z scores.

5.12 A Comparison of z Scores and Percentile Ranks

percentile rank the percentage of cases in a distribution that falls below a given point on the measurement scale (see Chapter 2)

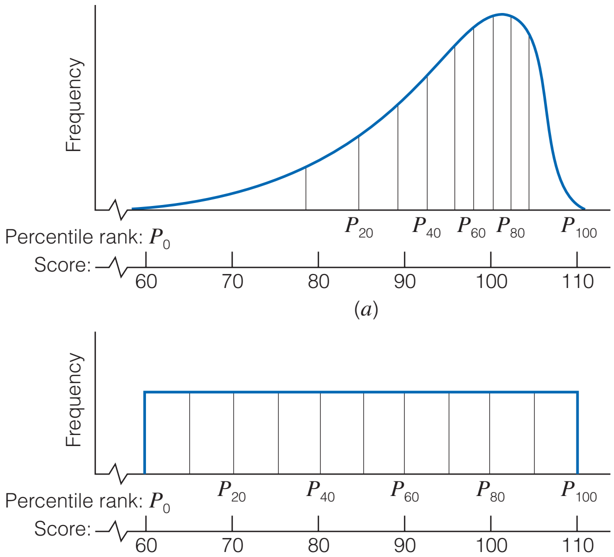
Similar to z scores, the **percentile rank** of a raw score describes its location relative to the other scores in the distribution (see Section 2.7). It has an advantage in terms of directness of meaning. The idea that Susan's percentile rank of 85 on a science test means that she performs better than 85% of comparable students is easy to understand and is meaningful even to most people without statistical training.

Unfortunately, there are some disadvantages. The primary problem is that *changes in raw scores are ordinarily not reflected by proportionate changes in percentile rank.* When one percentile is higher than another, the corresponding raw score of the one is higher than that of the other, but we do not know by how much. Look at Figure 5.9*a*, for example. A change of 10 points in percentile rank reflects a change in score that may be larger or smaller than 5 points, depending on the location in the distribution. Changes in raw score are accompanied by proportionate changes in percentile rank only when the distribution of scores is rectangular. The distribution in Figure 5.9*b* illustrates this. There, a change of 10 points in percentile rank consistently reflects a change of 5 points.

5.13 Summary

To describe a distribution adequately, we must provide a measure of variability as well as information about central tendency and shape. Measures of variability are summary figures that express quantitatively the extent to which the scores in a distribution scatter about or cluster together. We have considered four measures in this chapter: range, semi-interquartile range, variance, and standard deviation.

The range is the distance between the highest and lowest scores in the distribution. It is easy to obtain and useful for preliminary work, but as a descriptive measure it has the shortcoming of being responsive only to the location of the two most extreme scores.



(*a*)

(*b*)

FIGURE 5.9 Comparative location of percentile ranks in two distributions. The area under the curve in each segment equals 10% of the whole, representing 10% of the scores (see Section 3.4).

The semi-interquartile range is one-half the distance between the first (Q_1) and the third (Q_3) quartile points. Thus, it is related to the median (P_{50}) in many respects. It is responsive to the number of scores lying above or below the outer quartile points, but not to their exact location (magnitude). Because of this, the semi-interquartile range is particularly useful with open-ended distributions.

The variance is the mean of the squared deviation scores from the mean of the distribution: $S_X^2 = \Sigma(X - \overline{X})^2/n$. It is responsive to the exact position of each score in the distribution, but it is of little use in descriptive statistics because it is a quantity expressed in squared units. However, as you will see in later chapters, it has great importance in inferential statistics because of its resistance to sampling variation.

The standard deviation is the square root of the variance and therefore possesses all of its properties. Because it is expressed in original score units, it is the most widely used measure of variability in descriptive statistics. It is widely used in inferential statistics as well.

Measurements such as length and weight have a standard reference and thus have the same meaning to everyone. This is not true for most tests in the behavioral sciences. Without a frame of reference, raw scores on these tests are largely uninterpretable. We can compare variables with different measurement units by converting them to standard scores. One type of standard score is the *z* score. A *z* score states how far away a raw score is from the mean of the distribution using the standard deviation as the unit of measurement: $z = (X - \overline{X})/S_X$. When we change raw scores in a distribution to *z* scores, it does not change the shape of the distribution.

Mathematical Notes

Note 5.1 The Raw-Score Equivalent of $\Sigma(X - \overline{X})^2$ (*Ref.*: Section 5.5)

$$
\begin{aligned}
SS_X &= \sum(X - \overline{X})^2 \\
&= \sum(X^2 - 2X\overline{X} + \overline{X}^2) \\
&= \sum X^2 - 2\overline{X}\sum X + n\overline{X}^2
\end{aligned}
$$

$$= \sum X^2 - 2\left(\frac{\sum X}{n}\right)\left(\sum X\right) + n\left(\frac{\sum X}{n}\right)^2$$

$$= \sum X^2 - 2\frac{\left(\sum X\right)^2}{n} + \frac{\left(\sum X\right)^2}{n}$$

$$= \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

Note 5.2 The Standard Deviation as a Minimum Value When Deviations Are Taken from the Mean (*Ref.*: Section 5.8)

Let A be the point about which the deviation of each score is taken, and define it as a point that differs from the mean by an amount d: $A = \overline{X} + d$. Then,

$$\begin{aligned}\sum(X - A)^2 &= \sum[X - (\overline{X} + d)]^2\\ &= \sum[(X - \overline{X}) - d]^2\\ &= \sum[(X - \overline{X})^2 - 2d(X - \overline{X}) + d^2]\\ &= \sum(X - \overline{X})^2 - 2d\sum(X - \overline{X}) + nd^2\end{aligned}$$

From Note 4.1 at the end of Chapter 4, $\Sigma(X - \overline{X}) = 0$, so

$$\sum(X - A)^2 = \sum(X - \overline{X})^2 + nd^2$$

The right side of this equation shows that $\Sigma(X - A)^2$ is smallest when $d = 0$. From the definition of d, this occurs when $A = \overline{X}$. Because $S_X = \sqrt{\Sigma(X - \overline{X})^2/n}$, it is apparent that S has a minimum value when deviations are taken from $\overline{X}$.

Key Terms, Concepts, and Symbols

- measure of variability (p. 62)
- range (p. 63)
- semi-interquartile range (p. 63)
- Q (p. 63)
- quartile points (p. 63)
- Q_1, Q_2, Q_3 (p. 63)
- deviation score (p. 64)
- $X - \overline{X}$ (p. 64)
- variance (p. 65)
- σ^2 (p. 65)
- S^2 (p. 65)
- sum of squares (p. 65)
- SS (p. 65)
- $\Sigma(X - \overline{X})^2$ (p. 65)
- standard deviation (p. 66)
- σ (p. 66)
- S (p. 66)
- z score (p. 75)
- mean and standard deviation of z scores (p. 76)
- standard scores (p. 77)
- percentile rank (p. 77)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Which measure of variability—the range, semi-interquartile range, or the standard deviation—is (a) best for open-ended distributions? (b) calculated by using only two scores? (c) calculated by taking the sum of squared deviations from the mean? (d) not responsive to scores in the middle of the distribution? (e) best for very skewed distributions? (f) responsive to the exact position of every score in the distribution? (g) related to the median in its properties?

2. For the history midterm exam scores in Table 2.2, find (a) the range and (b) the semiinterquartile range.

3. We want to compare the variability of the two sets of reaction time scores in Data 3A. (a) Find the range of both distributions. (b) Find Q for both distributions. (c) Which set of scores is more variable?

4. On the midterm exam in their anthropology course, Jan's deviation score was +5, Kim's was −12, and Leslie's was 0. (a) Of these three, who did the best and who did the worst? (b) Exactly where did each fall in the distribution in relation to the mean?

5. On the exam cited in Problem 4, the mean for the class was 80. What was the raw score for (a) Jan, (b) Kim, and (c) Leslie?

6. A distribution consists of five scores. For four of the scores, the deviations from the mean are +5, +2, +1, and −8. What is the deviation of the fifth score?

7. The mean of the distribution cited in Problem 6 is 10. What are the raw scores?

8. For the nine scores shown in Figure 4.2, find (a) the variance and (b) the standard deviation. Use the deviation-score method.

9. Use the raw-score method to calculate the variance and standard deviation for the same nine scores in Problem 8.

10. Distribution X is 5.1, 8.7, 3.5, 5.4, and 7.9. (a) Find the range. Using the raw-score method and keeping all work to one decimal place, find (b) the variance and (c) the standard deviation. Using the deviation-score method, find (d) the variance and (e) the standard deviation. (f) Which method is easier?

11. If the scores in Figure 4.2 are the lengths in inches of rats' tails, what is the unit of measurement for (a) their variance and (b) their standard deviation?

12. A ninth-grade science teacher gives a standard achievement test to his class and finds that $\overline{X} = 90$ with $S_X = 8$. For a national sample of ninth graders, $\overline{X} = 75$ with $S_X = 14$. (a) How do his students compare to the national sample? (b) What do the data suggest for teaching science to this class?

13. Distribution X is: 15, 14, 11, 11, 9, and 6. Using the deviation-score method, find (a) the variance and (b) the standard deviation. Distribution Y is: 17, 16, 13, 13, 11, and 8. Using the deviation-score method, find (c) the variance and (d) the standard deviation. By both measures, X and Y have the same variability. (e) Why does the variance work out to be the same for both distributions? (f) Why does the standard deviation work out to be the same for both?

14. In 88 infants, researchers measure the change in heart rate before and after brightening the room. The distribution of changes ranges from −20 to +5 beats per minute, with a mean of −9.7. (Heart rate usually slows when an organism pays attention to a stimulus.) Upon rechecking, the lowest score turns out to be −24. Which measures of variability are affected by the correction?

15. (a) Can the distribution of heart-rate changes cited in Problem 14 have a variance of −16? Explain. (b) Can the distribution have a standard deviation of −4? Explain.

16. Seven scores are 2, 5, 7, 8, 9, 11, and 14. Find (a) S and (b) Q. Now add the extreme score of 24. For the modified distribution (now with $n = 8$), find (c) S and (d) Q. (e) By what percentage did S increase? (f) By what percentage did Q increase? (g) What generalization is illustrated in comparing answers (e) and (f)?

17. In reporting the frequency distribution of annual household income, the *Statistical Abstract of the United States* uses these class intervals: under $10,000; $10,000–14,999; $15,000–24,999; $25,000–34,999; $35,000–49,999; $50,000–74,999; and $75,000 and over. Which measures of variability could we compute for the distribution as given in terms of those class intervals?

18. The standard deviation of a set of scores is 20. What will the standard deviation become if (a) 15 points are added to each score? (b) Each score is divided by 5? (c) 10 points are subtracted from each score and the resulting values are divided by 2? (d) Each score is multiplied by 5, and 20 is then added to each resulting value?

19. A consumer testing company evaluates eight different brands of tires. The tires are tested individually in a laboratory on an artificial surface to control for differences in driving habits and road conditions. The number of miles elapsed until 2/16 inch of tread remains is recorded. The results are 45,000; 40,000; 80,000; 70,000; 65,000; 60,000; 60,000; and 50,000. Find (a) the range and (b) the standard deviation.

20. When answering Problem 19, did you calculate the standard deviation by using the numbers that were given (e.g., 45,000)? If so, how could you have transformed the scores to make the calculation less cumbersome (if you had to calculate by hand)?

21. The Wechsler Intelligence Scale has a mean of 100 and a standard deviation of 15. Find the z score (to two decimal places) of an IQ score of (a) 115, (b) 130, (c) 70, (d) 100, (e) 80, (f) 95, (g) 108, (h) 87, and (i) 122.

22. For the same intelligence test, find the actual IQ score for the following z score equivalents (use Formula 5.5): (a) -0.60, (b) $+1.40$, (c) $+2.20$, (d) -1.80, (e) 0.00, (f) -2.40, (g) $+0.20$, and (h) $+4.40$.

23. A student's grade on a very difficult English exam is 85. What can we say about the merits of the student's performance? Explain.

24. The mean of a set of z scores is always zero. Does this mean that half of a set of z scores will be negative and half positive? Explain.

25. Susan and Robin are arguing about who did better on their final exams. Susan got an 88 on her psychology final exam (mean = 74, standard deviation = 12), whereas Robin earned an 82 on her biology exam (mean = 76, standard deviation = 4). All freshmen are required to take both classes. Who did better? Explain.

26. Bob and Bert are also debating about who did better on their exams. Bob got a 65 on the psychology exam, and Bert got a 68 on the biology exam. Who did better? Explain.

27. For Problem 19, find the z scores of the eight scores. Has the shape of the distribution of eight scores changed? Explain.

28. It will help you learn how to apply the formulas for the variance and standard deviation if you work many problems. You do not need large data sets. If you think you need to work more problems, find the variance and standard deviation by both formulas for the small data sets in Problem 1a–i in Chapter 4.

29. For the data in Table 2.1, use IBM SPSS to calculate the mean and standard deviation. (Refer to Section 5.6 to adjust the value for the standard deviation provided by IBM SPSS so that the *SS* are divided by n.)

30. For the Problems in Chapter 2, use IBM SPSS to calculate the mean and standard deviation on (a) Data 2A in Problem 8, (b) Data 2C in Problem 18, and (c) Data 2D in Problem 19. (Refer to Section 5.6 to adjust the value for the standard deviation provided by IBM SPSS so that the *SS* are divided by n.)

CHAPTER 6

Standard Scores and the Normal Curve

When you have finished studying this chapter, you should be able to:

- Summarize the history of the concept of the normal curve;
- Know the defining characteristics of a normal curve;
- Use the normal curve table to find areas under the normal curve that fall above or below a specified score;
- Use the normal curve table to find a score when the area (under the normal curve) above or below the score is known; and
- Understand that although no real variable is exactly normally distributed, the normal curve serves as a useful model for many real variables and also for distributions of sample means.

Suppose we toss 8 coins and record the number of heads. It can vary from 0 to 8. Each coin is as likely to come up heads as tails, and the outcome of the toss of any one of the 8 coins in no way affects the outcome of the toss of any other. Because of chance factors, we cannot say precisely how many heads will appear in any one trial of 8 coins. But suppose we ask other

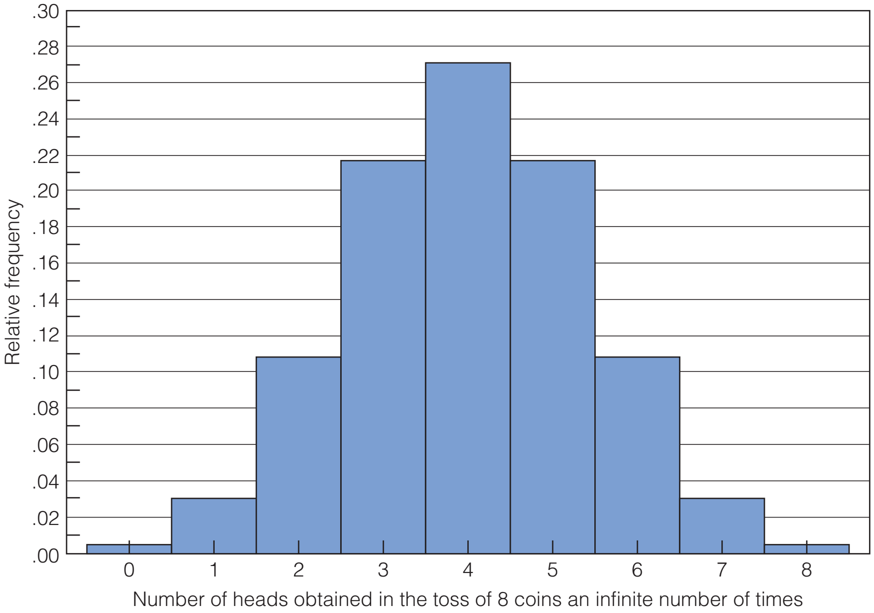


FIGURE 6.1 Relative frequencies of heads obtained when 8 coins are tossed an infinite number of times.

questions: If we repeat the act of tossing the 8 coins an extremely large number of times, how often will we obtain 0 heads? How often will heads come up once? Twice? Early mathematicians developed a theory that provides the answers. We show the relative frequency distribution derived from that theory (called a *binomial distribution*) in Figure 6.1. The figure shows that *on the average* there will be 4 heads and that the more divergent the number of heads and tails in any one trial (e.g., 7 heads and 1 tail), the lower the relative frequency of its occurrence.

If we superimpose a curve on the histogram in Figure 6.1, it approximates the bell-shaped, normal curve shown in Figure 3.13*f*. If we tossed 16 coins, rather than 8, there would be more bars in the histogram, and the normal curve would be an even better fit, as illustrated in Figure 6.2. If we toss an even greater number of coins, the number of bars in the histogram similarly increases, and the histogram grows smoother. It is obvious that in this situation, where the outcome of events is generated by the operation of chance factors, the normal curve gives us a good approximation to the actual relative frequency distribution. Before we proceed, some history of the normal curve may help you understand its widespread use as a model for statistical observations.

6.1 Historical Aspects of the Normal Curve

In 1733, the French mathematician **Abraham de Moivre** discovered the formula for the normal curve as the limiting case characterizing an infinite number of independent events (such as coin tosses). His work was initially ignored, and further significant development did not occur until the beginning of the 19th century. Then, it seems that two brilliant mathematicians, the Frenchman **Pierre-Simon** (the Marquis de Laplace) and the German **Carl Friedrich Gauss**, rediscovered the normal curve independently. Gauss was interested in

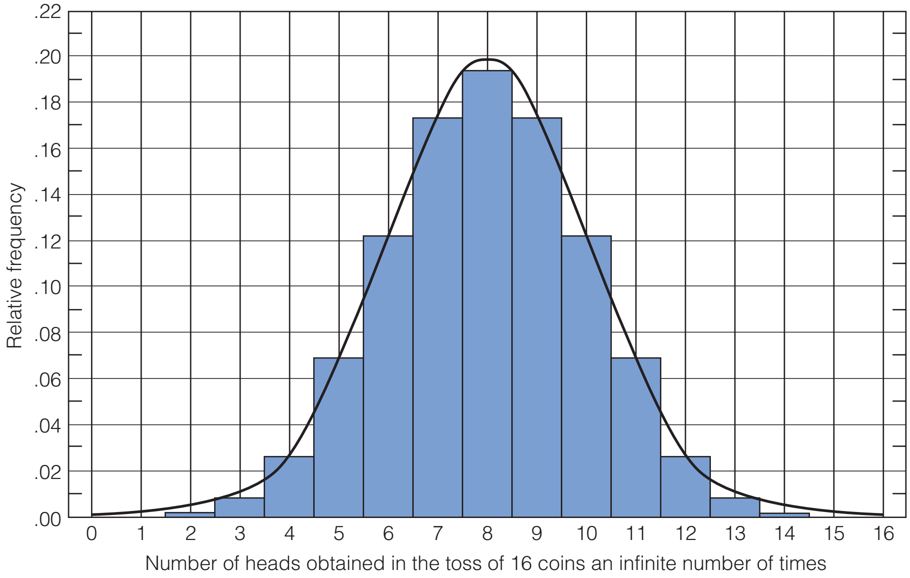


FIGURE 6.2 Relative frequencies of heads obtained when 16 coins are tossed an infinite number of times.

problems in astronomy, such as determining the orbit of a planet. His solutions were based on observations that were subject to error (which he found to be distributed in a bell-shaped fashion), and this led him to a theory of errors of observation. It had such strength that in the 1800s the normal curve came to be known as the "normal law of error." The name is now obsolete, but vestiges of this terminology hang on. In Chapter 11, for example, you will learn about the term *standard error.*

The use of the normal curve as a model for other situations is due primarily to the work of **Adolphe Quetelet**, a Belgian, in the middle part of the 19th century. He taught mathematics and astronomy, created the basic methods of physical anthropometrics, and helped found sociology. He believed that the normal curve could be extended to apply to problems in meteorology, anthropology, and human behavior. It was his opinion that "mental and moral traits," when measured, would conform to the "normal law." One of his conceptions was *l'homme moyen* or "the average man." He compared the normal distribution to the notion that Nature, in aiming at a mark (*l'homme moyen*), missed and so produced a distribution of "error" in her effort.

In the late 19th century, the British genius **Sir Francis Galton** began the first serious study of individual differences, an important topic for the behavioral sciences. He found that many mental and physical traits conformed reasonably well to the normal curve, and he was greatly impressed by the applicability of the normal curve to natural phenomena. In 1889 he wrote:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. F. GALTON, *Natural Inheritance,* MacMillan and Co., London, 1889, p. 66.

His enthusiasm is understandable and worthy of our appreciation. His theory, however, must be tempered in light of current knowledge. We will present a more balanced view in the coming sections.

So the time has come for us to examine the normal curve more closely. We need to know what the normal curve is, what its properties are, the ways in which it is useful as a statistical model, and how to put it to work in answering questions.

6.2 The Nature of the Normal Curve

normal curve a mathematical abstraction with a particular defining equation (see Mathematical Note 6.1); the equation describes a family of normal curves

characteristics of normal curves symmetrical; unimodal; asymptotic with the horizontal axis; a continuous distribution; the proportion of the area under the curve relative to a particular range of values on the horizontal axis is the same when the values are stated on a comparable basis

It is important that you learn to distinguish between data that are normally distributed (or approximately so) and the normal curve itself. The **normal curve** is a mathematical abstraction with a particular defining equation (given in Note 6.1 at the end of this chapter). As a mathematical abstraction, it is not associated with (or determined by) any event or events in the real world. It is not, therefore, a "law of nature," contrary to the thought and terminology associated with it in the 1800s. In the same way, the equation of a circle is a mathematical abstraction and not a law of nature.

Just as the equation of a circle describes a family of circles, some big and some small, so *the equation of the normal curve describes a family of normal curves.* The family differs in their means and standard deviations. We illustrate three possibilities in Figure 6.3: equal means and unequal standard deviations, unequal means and equal standard deviations, and unequal means and unequal standard deviations.[1]

In other respects, members of this family have the same characteristics. All normal curves are **symmetrical distributions**; that is, the left half of the normal curve is a mirror image of the right half. They are **unimodel distributions**, with the mode at the center. Indeed, the mean, median, and mode all have the same value. If we start at the center of the curve and work outward, we see that the height of the curve descends gradually at first, then faster, and finally slower. A curious and important situation exists at the extremes of the curve. Although the curve descends quickly toward the horizontal axis, it theoretically never actually touches it, no matter how far out one goes. It is therefore said to be an **asymptotic curve** with respect to the horizontal axis. You might think that if the curve never touches the X axis, a large proportion of the total area under the curve must lie in the extremes, but this is not so. For example, the area included within the range of $\mu \pm 3\sigma$ comprises approximately 99.7% of the total area (see Figure 5.7).

The normal curve is a **continuous distribution**. However, remember that the process of measurement always reduces a continuous variable to a discrete variable (see Section 1.4). Thus, when the normal curve is used as a model for events of the real world, it can be only an approximation.

Finally, *for all normal curves, the proportion of the area under the curve relative to a particular range of values on the horizontal axis is the same when the values are stated on a comparable basis*. For example, approximately two-thirds (68.26%) of the area under a normal curve lies within the range of $\mu \pm 1\sigma$ (see Figure 5.3). Thus in a normal curve characterized by $\mu = 100$ and $\sigma = 20$, about two-thirds of the area falls between the values of 80 and 120, and in a normal curve characterized by $\mu = 500$ and $\sigma = 100$, about two-thirds of the area falls between the values of 400 and 600. We will consider problems of this type next.

---

[1] When the normal curve is an appropriate model for real events, it is most often the population of such events that is so modeled. Consequently, we shall use μ and σ in connection with the normal curve throughout this chapter rather than $\overline{X}$ and S.

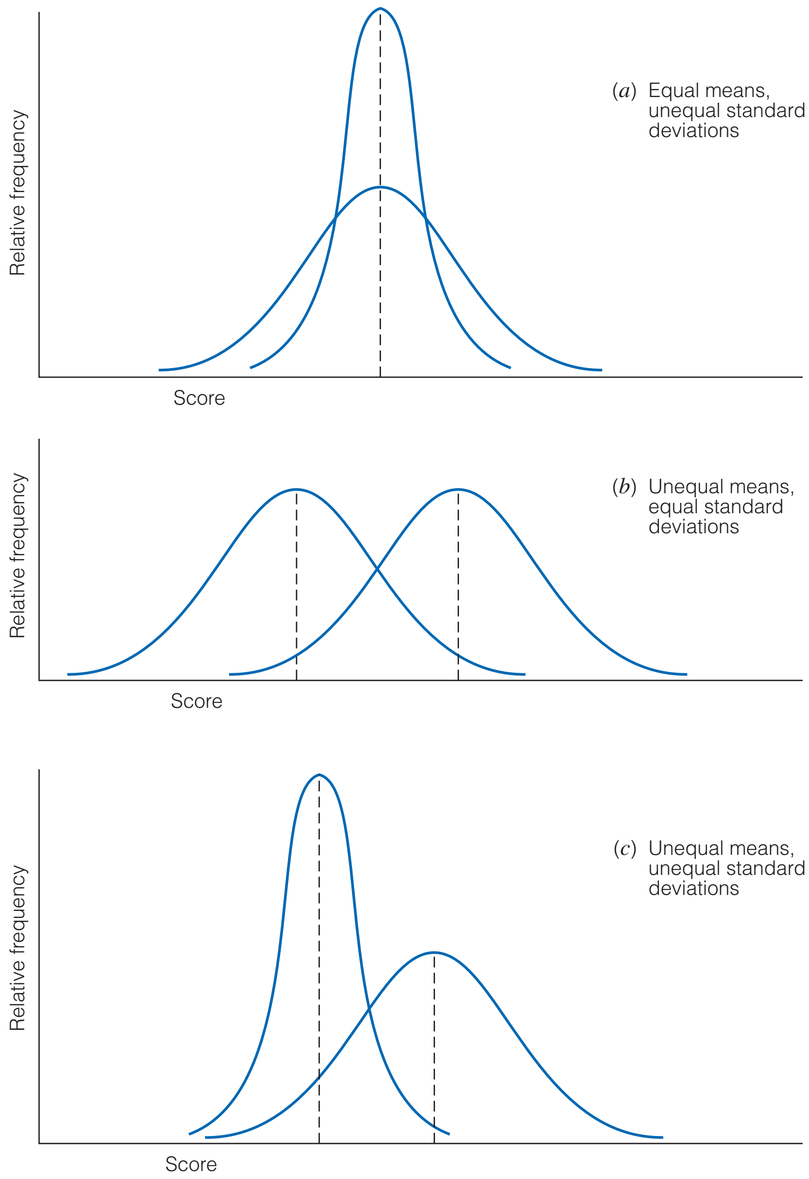


FIGURE 6.3 Variations of normal curves in means and standard deviations.

6.3 Standard Scores and the Normal Curve

Recall that the equation defining the normal curve actually describes a whole family of curves. Their means vary, and their standard deviations vary; it is their *shape* that they have in common. To use the normal curve as a model, we need a way to locate a position within a normal distribution that does not depend on the particular mean and the particular standard deviation of the distribution. We can do this by converting raw scores to standard scores.

Look at Figure 6.3. In what sense do the curves there all have the same shape, the shape called normal? *The shapes are the same in that the area under each curve is distributed in the same way.* Thus, we need only a single table of the areas under the normal curve for work with *all* normal distributions, irrespective of their means and standard deviations. All that we need to do is translate the relevant raw score to a z score in order to use the table. Table A in Appendix D

"I don't know anything about the bell curve,
but I say heredity is everything."
28320

The Cartoon Bank (2004) Psychology Catalog.

gives you the area under the upper half of the normal curve that falls above or below a particular point along the horizontal axis when that point is *expressed as a standard score.*

standardized normal curve a normal distribution in which scores have been transformed to **standard** (*z*) **scores**

Why is it important that you learn to calculate area under the normal curve for a specified span along the abscissa? Because *the proportion of area under the curve corresponds to the proportion of scores in the distribution.* Thus, if about 68% of the area under the normal curve falls within 1 standard deviation of the mean (above *and* below), we know that about 68% of the cases in the distribution fall within that range. For example, if IQ as measured by the Wechsler Adult Intelligence Scale (WAIS) is normally distributed with a mean of 100 and standard deviation of 15, we know that about 95% of adults in the United States have an IQ between 70 and 130. The next two sections are concerned only with distributions that follow the normal curve.

6.4 The Standard Normal Curve: Finding Areas When the Score Is Known

Given a distribution of scores, what proportion of the scores lie above a score of 120? Between 85 and 120? If the shape of the distribution closely resembles the normal curve, we can use Table A in Appendix D to answer questions like these. In this table, the total area under the curve is defined as unity. To use the table, remember that the normal curve is symmetrical and therefore has 50% of its total area below the mean ($z = 0$) and 50% of its area above the mean.

Case 1. Finding the Area under the Normal Curve that Falls above a Known Score

You will frequently see questions of the following type in statistical work: Given a normally distributed variable with a mean of 100 and a standard deviation of 20, what proportion of scores will fall above 120? Suppose, for example, that 3,000 entering freshmen at State University are given an entrance examination in mathematics and that the scores are distributed normally with

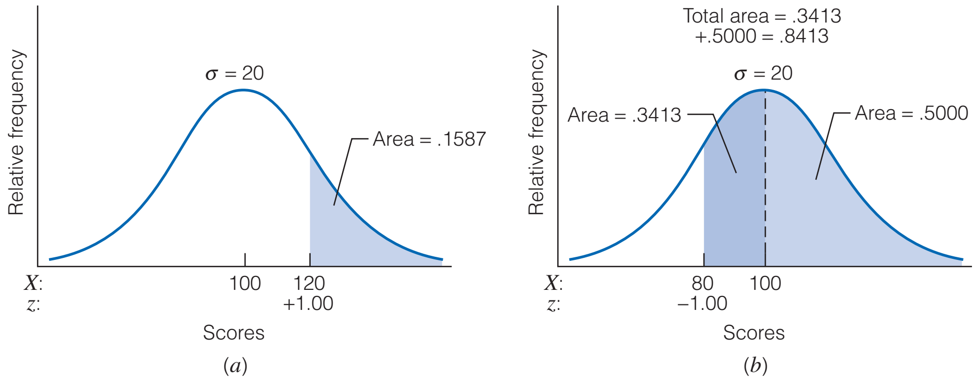


FIGURE 6.4 Proportion of scores in a normal distribution characterized by $\mu = 100$ and $\sigma = 20$ (*a*) exceeding a score of 120 and (*b*) falling above a score of 80.

a mean of 100 and a standard deviation of 20. If the university decides to place all students scoring above 120 into honors math, how many students will be placed into the class? We illustrate this problem in Figure 6.4*a*.

The first step is to translate the score of 120 to a z score: $(120 - 100)/20 = +1.00$. Now turn to Table A in Appendix D. Locate the value $z = 1.00$ in the first column and look across that row for the entry in the column headed "Area beyond z." This entry is .1587, which tells us that in *any* batch of scores that is normally distributed, .1587 of the scores fall above $z = +1.0$, regardless of the values of μ and σ. Because the total area under the curve is defined as unity in this table, .1587 is the proportion of the total area falling in the tail. As we noted in the last section, area under the curve is proportional to the frequency of scores. Thus, our question is answered as follows: $(.1587 \times 3{,}000)$ or 476 freshmen will be placed in honors math.

Suppose the problem had been to determine the number of freshmen scoring above 80 on the exam. Figure 6.4*b* illustrates this problem. The z-score equivalent of 80 is -1.00. Negative z scores do not appear in the table. Because the normal curve is symmetrical, as many scores fall below a z score of -1.00 as fall above a z score of $+1.00$. The table therefore does not need to show both positive and negative values of the z scores. In this problem, we already know that .5000 of the scores fall above the mean, so all we need to determine is the proportion of scores between the mean and a score that is 1 standard deviation below the mean. You can find this information in Column 2 of Table A. For a score with a z value of -1.00, that proportion is .3413. Adding .3413 to .5000, we find the total proportion of students above 80 to be .8413. Thus, about $(.8413 \times 3{,}000)$ or 2,524 students scored above 80 on the entrance exam.

Case 2. Finding the Area under the Normal Curve that Falls below a Known Score

Suppose that State University has also decided to place all students scoring less than 85 into remedial math. How many must take the remedial course? This problem is illustrated in Figure 6.5*a*. To solve the problem, we first determine the z-score equivalent of 85, which is $(85 - 100)/20 = -15/20 = -.75$. We use Column 3 in Table A and see that in any normal distribution, .2266 of the area falls below a score that is three-quarters of a standard deviation below the mean. Thus, $.2266 \times 3{,}000 = 679.8$, or roughly 680 freshmen will have to take remedial math.

As you can see, answers to this type of problem are found very much like those for the first. For example, how many students at State University were not placed into honors math (how many scored below 120)? This is illustrated in Figure 6.5*b*. We know that .5000 of the scores fall below the mean, so we must add this to the proportion of scores falling between the mean and a score 1 standard deviation above the mean: $z = (120 - 100)/20 = +1.00$. Column 2

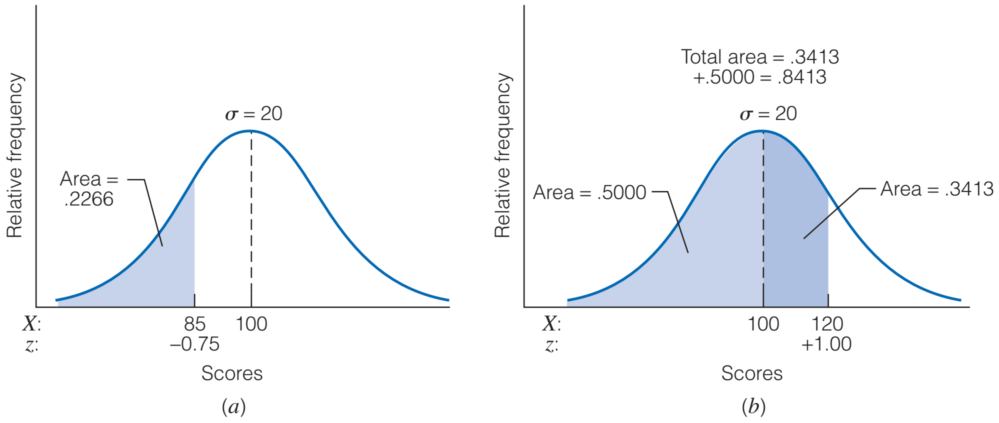


FIGURE 6.5 Proportion of scores in a normal distribution characterized by $\mu = 100$ and $\sigma = 20$ (*a*) below a score of 85 and (*b*) below a score of 120.

shows that the proportion is .3413, which when added to .5000 gives .8413. Thus, (.8413 × 3,000) or about 2,524 freshmen did not earn a high enough score on the entrance exam to be placed in honors math.

Case 3. Finding the Area under the Normal Curve that Falls between Two Known Scores

Given the entrance examination results of Case 1, how many entering freshmen at State University were placed into regular freshmen-level math courses (i.e., were not placed into either remedial or honors courses)? In other words, how many scored between 85 and 120? This problem appears in Figure 6.6*a*. Again, the first step is to obtain the z-score equivalents of these two scores. The values are $z = -.75$ and $z = +1.00$, respectively. Find the area under the curve between the z score of $-.75$ and the mean and add this to the area between the mean and the z score of $+1.00$. In Table A, the second column gives the appropriate values:

Area between $z = -.75$ and the mean:	.2734
Area between the mean and $z = +1.00$:	.3413
Total area:	.6147

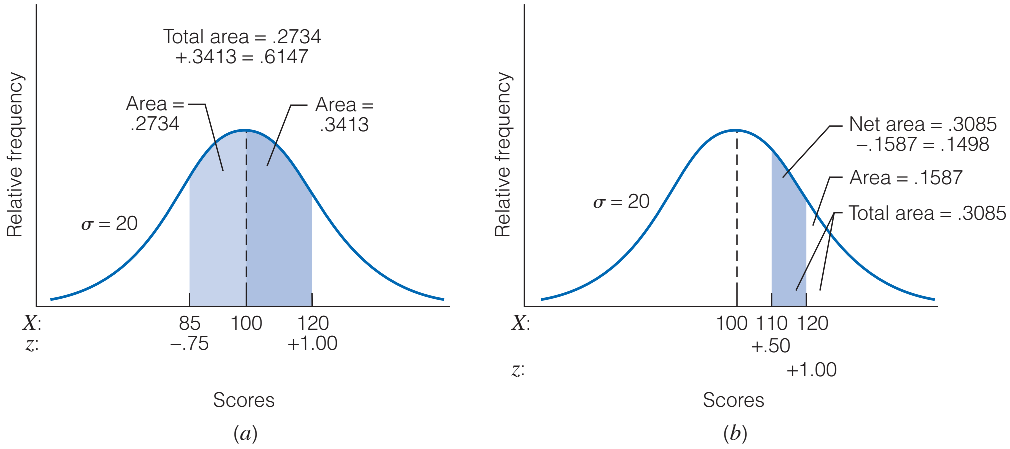


FIGURE 6.6 Proportion of scores in a normal distribution characterized by $\mu = 100$ and $\sigma = 20$ falling (*a*) between scores of 85 and 120 and (*b*) between scores of 110 and 120.

Multiplying 3,000 by .6147, we find that 1,844.1 or roughly 1,844 students scored between 85 and 120 on the entrance exam.

For the same distribution of entering freshmen, suppose we wanted to know how many scores fall between the values of 110 and 120. One way to answer is to determine the proportion of scores falling above 110 and subtract from this value the proportion of scores falling above 120. The problem and its solution appear in Figure 6.6*b*. We begin, as usual, by converting the raw scores to z scores: $X = 110$ becomes $z = +.50$ and $X = 120$ becomes $z = +1.00$. With use of Table A, the solution is:

$$\begin{array}{ll} \text{Area above } z = +.50\text{:} & .3085 \\ \text{Area above } z = +1.00\text{:} & -\underline{.1587} \\ \text{Difference:} & .1498 \end{array}$$

We multiply .1498 × 3,000 and get 449.4 or roughly 449 scores.

6.5 The Standard Normal Curve: Finding Scores When the Area Is Known

In the previous section, we knew the score and wanted to find the area under the curve. Here, we know the area and want to find the score. We solve these problems by reading Table A in Appendix D backward, that is, by locating the area in Column 2 or 3 and then identifying the z-score equivalent of the desired value.

Case 1. Finding the Score above or below Which a Certain Percentage of the Total Scores Fall

Suppose that neighboring City University also has remedial and honors math courses for entering freshmen, but it determines placement in a slightly different manner. Students take the same entrance exam, but City University places the top 20% into honors math. Let us assume that the mean is again 100 and the standard deviation is again 20. What score separates the upper 20% of the entering freshmen from the lower 80%? If (and only if) the shape of the distribution of scores is well approximated by the normal curve, Table A can help us.

The problem appears in Figure 6.7*a*. In Table A, we look for a value as close as possible to .2000 in the column headed "Area beyond z" and find that the z score associated with it is 0.84. The 0.84 can be positive or negative. Which do we want? Because we want to separate the top 20% from the remainder, the value of the z score is positive: $z = +0.84$. We then substitute this value into our formula for z and solve for X:

$$z = \frac{X - \mu_X}{\sigma_X}$$

$$+0.84 = \frac{X - 100}{20}$$

$$(.84)(20) = X - 100$$

$$16.8 = X - 100$$

$$116.8 = X$$

Thus, to be placed into honors math at City University, a freshman would have to score at least 117 on the entrance exam.

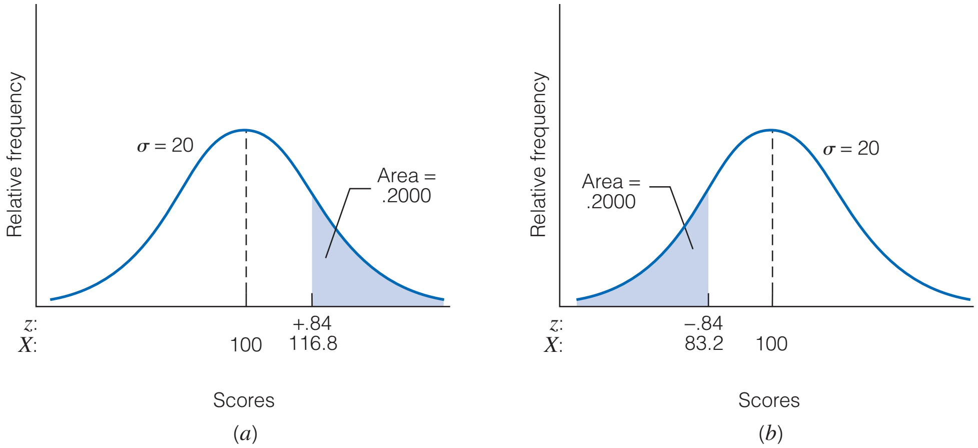


FIGURE 6.7 The score dividing (*a*) the upper 20% and (*b*) the lower 20% from the remainder of scores in a normal distribution characterized by $\mu = 100$ and $\sigma = 20$.

We worked the previous problem by starting with the formula for a z score and solving for X. You may find it easier to solve for X by the following equation, which is derived from the z-score formula:

$$X = \mu_X + z\,(\sigma_X)$$

If City University places students who scored in the bottom 20% on the exam in a remedial math course, what score separates these students from the upper 80%? This problem is illustrated in Figure 6.7*b* and we work it just like before except that, in this case, we must use the negative value for the z score associated with .2000 in Column 3 of Table A. When we substitute this value, $-.84$, in the formula for z, we find that $X = 100 - 16.8$, or 83.2. Thus, students who scored 83 or below on the exam would have to take remedial math.

Case 2. Finding the Limits within Which a Certain Percentage of Scores Fall Equidistant from the Mean

Suppose City University is interested in knowing the values on the mathematics entrance examination within which the central 90% of the scores fall. The problem appears in Figure 6.8. In this case, 45% of the scores within the range of interest fall above the mean

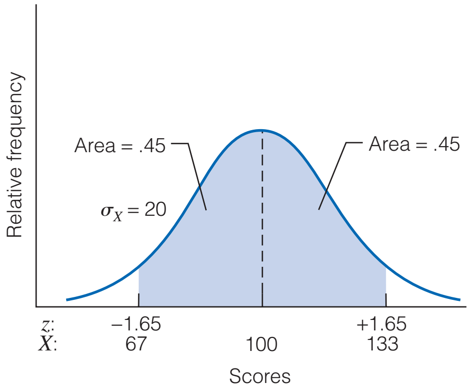


FIGURE 6.8 The limits that include the central 90% of scores in a normal distribution characterized by $\mu = 100$ and $\sigma = 20$.

and 45% fall below the mean. What is the score that has 45% of the scores between it and the mean? From Table A, we find that it has a z value of 1.65. The scores are therefore located 1.65 standard deviations above and below the mean. We substitute this value into the formula for z and solve for X:

LOWER LIMIT	UPPER LIMIT
$z = \dfrac{X - \mu_X}{\sigma_X}$	$z = \dfrac{X - \mu_X}{\sigma_X}$
$-1.65 = \dfrac{X - 100}{20}$	$+1.65 = \dfrac{X - 100}{20}$
$(-1.65)(20) = X - 100$	$(+1.65)(20) = X - 100$
$-33 = X - 100$	$+33 = X - 100$
$67 = X$	$133 = X$

Thus, 90% of the scores on this exam fall between $(100 - 33) = 67$ and $(100 + 33) = 133$.

6.6 The Normal Curve as a Model for Real Variables

We now know that the normal curve is a rather good fit to a large number of distributions of real data in a wide variety of sciences. The distribution has applications in the fields of physics, chemistry, biology, engineering, and economics, to name only a few. For example, in reasonably *homogeneous* human populations, the distribution of height follows the normal curve closely. Thus, the normal curve offers a convenient and reasonably accurate *description* of a great number of distributions.

Several words of caution are in order, however. First, the normal curve as a model best describes an infinite number of observations that are on a continuous scale of measurement. As we know, recorded observations are discrete rather than continuous, and among concrete data we do *not* have an infinity of observations. Real populations may be large, but they are not of infinite size. Although the normal curve closely approximates distributions of data of many kinds, it is a fair speculation that *no real variable is exactly normally distributed*. Second, *many variables are not normally distributed*. This includes the results of many tests in the behavioral sciences (see Point of Controversy).

6.7 The Normal Curve as a Model for Sampling Distributions

The second way in which the normal curve functions as a model is for a distribution of sample statistics. For instance, if we draw a great number of random samples from a population, each sample of the same size, and compute the mean of each sample, the distribution of this large number of means will approximate the normal curve. This is a property of utmost importance in statistical inference because, as you will see in later chapters, we must know the shape of the distribution of sample means to make an inference about the population mean.

We will see the normal curve in its role as a model from time to time in the next few chapters. Then, beginning with Chapter 11 and throughout most of the remaining chapters, it will be of central importance.

Point of Controversy

How Normal Is the Normal Curve?

Galton (1889) believed that so many mental and physical traits conformed to the bell-shaped normal curve that it was an expression of cosmic order (the "Law of Frequency of Error"). Though statisticians today do not regard the normal curve as a law, many have continued to assume that real-world variables are frequently distributed in this manner. One textbook even joked that "God loves the normal curve!" (Hopkins, Glass, & Hopkins, 1987, p. 51).

Doubts concerning the usefulness of the normal curve as a model for the distribution of real-world variables occasionally surfaced. Nunnally (1978), for example, suggested that because tests of mental ability are usually constructed with items that are positively interrelated, the "test scores are seldom normally distributed." Others also expressed their doubts: "Frequently a few large errors infect the data so that the tails of the underlying distribution are heavier than those of the Gaussian distribution" (Hoaglin, Mosteller, & Tukey, 2000). Geary (1947) earlier voiced even stronger doubts: "Normality is a myth; there never was, and never will be, a normal distribution."

What the naysayers lacked, however, was empirical data to support their claim that large discrepancies from normality occur frequently in behavioral science measures—until Theodore Micceri, that is. Micceri (1989), a professor of education at the University of South Florida, examined the distributions of 440 measures of achievement and psychological traits obtained in applied settings. Most of the distributions came from recently published research articles, and the remainder from national tests, tests used statewide in Florida, districtwide tests, and college entrance and Graduate Record Examination scores at the University of South Florida. Nearly 70% included samples of 1,000 or more. Only 19 of the 440 distributions were found to approximate the normal curve. Of the remainder, only half of the distributions could be defined as relatively smooth; the others were "lumpy" (29% were bimodal or multimodal). Half of them had at least one extremely heavy tail, and nearly one-third were found to be extremely asymmetric.

The normal curve has been a useful mathematical model for statisticians, particularly because of their faith in the central limit theorem (see Section 11.4). This states that the distribution of an infinite number of sample means (from samples of a fixed size randomly drawn from a population) approaches normality as the sample size increases, irrespective of the shape of the population of scores. The greater the population differs from normality, the greater the sample size must be for the central limit theorem to hold. As you will learn in later chapters, we use this theoretical distribution of sample means to make inferences about the mean of the population.

Research regarding the effects of testing samples drawn from nonnormal distributions has generally looked at smooth mathematical models with tails greater than or less than the normal curve's. This research has found that the commonly employed tests used in inferential statistics are good under these nonnormal population conditions. However, until recently no one had yet examined how well these tests hold up using population distributions that vary from the normal curve as much as the "lumpy" and asymmetric real-world distributions found by Micceri. Although the implications of Micceri's findings for inferential statistics are still somewhat debated, as noted by Wilcox (1998, 2005), it appears that many commonly used statistical tests may not be as good as previously thought when using small samples and assumptions of normality are violated. (See the Point of Controversy in Chapter 21 for related material.)

6.8 Summary

The normal curve is a mathematical abstraction defined by an equation. The equation describes not just one curve, but a family of them. Although they may differ with respect to means and standard deviations, all normal curves are symmetrical, unimodal, and continuous, with tails that are asymptotic to the horizontal axis.

The area under all normal curves is distributed in the same way. Thus, to use the normal curve as a model, we need a way to locate a position within a normal distribution that does not depend on the particular mean and standard deviation of the distribution. We can compare variables with different measurement units by converting them to z scores. Changing raw scores in a distribution to z scores does not change the shape of the distribution. Thus, when we convert raw scores in a normal distribution to z scores, the result is the standard normal curve. Table A in Appendix D gives the proportion of the area under the upper half of the curve that falls beyond a stated value of z or between the mean and a particular value of z. Because the area under the curve is proportional to the frequency of scores in a distribution, we can use this table to calculate the number of scores in a normal distribution that fall above or below any value or between any two score points. Given a proportion of the scores (for example, the upper 20%), we can also determine the value that separates this proportion from the remainder of the distribution.

There is no real variable that is exactly normally distributed, but the resemblance is close enough in some cases that the normal curve serves as a useful model. As we will see in later chapters, this is especially true for distributions of some sample statistics (such as $\overline{X}$).

Mathematical Note

Note 6.1 The Equation of the Normal Curve (*Ref.*: Section 6.2)

The equation of the normal curve is

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

As a model for a frequency distribution, it may be interpreted as follows:

where Y = height of the curve directly above any given value of X

N = number of cases

X = a raw score

μ = mean of the set of scores

σ = standard deviation of the set of scores

$$\left.\begin{array}{r} \pi = 3.1416 \\ e = 2.7183 \end{array}\right\} \text{mathematical constants}$$

The variables that determine what Y will be are X, μ, σ, and N. Basically, the area under the curve is determined by N, the location of the center of the curve by μ, and the rapidity with which the curve approaches the abscissa by σ.

For a normal curve of unit area, in which the scores are standard scores, $N = 1$, $\mu = 0$, and $\sigma = 1$, and the equation is

$$Y = \frac{N}{\sqrt{2\pi}} e^{-(z^2/2)}$$

In this form, the value of z is the only variable that determines Y, the relative frequency. Strictly speaking, Y in the normal curve is not interpretable as frequency; it is simply the height of the curve corresponding to a particular (point) value of X. Frequency *is* interpretable as the area under the normal curve that falls between two values of X (the frequency of scores within that range).

Key Terms, Concepts, and Symbols

Abraham de Moivre (p. 83)
Carl Friedrich Gauss (p. 83)
Adolphe Quetelet (p. 84)
Sir Francis Galton (p. 84)
normal curve (p. 85)
symmetrical distribution (p. 85)
unimodal distribution (p. 85)
asymptotic curve (p. 85)
continuous distribution (p. 85)
standardized normal curve (p. 87)
standard score (p. 87)
z score (p. 87)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In the population of adult Americans, the variable IQ is found to be normally distributed. Discuss the similarities and differences between a normally distributed variable and the normal curve.

2. For normally distributed scores, what proportion of scores would fall: (a) above $z = +1.00$? (b) above $z = +2.00$? (c) above $z = +3.00$? (d) below $z = -2.00$? (e) below $z = -3.00$?

3. For normally distributed scores, what proportion of scores would fall: (a) above $z = -1.00$? (b) below $z = 0$? (c) below $z = +2.00$? (d) below $z = +0.85$?

4. For normally distributed scores, what proportion of scores would fall: (a) between $z = -1.00$ and $z = +1.00$? (b) between $z = -2.00$ and $z = +2.00$? (c) between $z = -3.00$ and $z = +3.00$? (d) between $z = -0.75$ and $z = +1.25$?

5. For normally distributed scores, what proportion of scores would fall: (a) between $z = +.25$ and $z = +1.25$? (b) between $z = -1.00$ and $z = -2.00$? (c) between $z = +2.00$ and $z = +3.00$?

6. For normally distributed scores, what proportion of scores would fall: (a) outside the limits $z = -2.00$ and $z = +2.00$? (b) outside the limits $z = -3.00$ and $z = +3.00$? (c) outside the limits $z = -1.00$ and $z = +2.00$?

7. If scores are normally distributed with a mean of 500 and a standard deviation of 100, what proportion of the scores fall: (a) above 550? (b) above 600? (c) below 420? (d) above 350? (e) between 550 and 750? (f) between 300 and 625? (g) between 625 and 725? (h) between 380 and 480?

8. In a normal distribution of 1,000 aptitude test scores, with a mean of 60 and a standard deviation of 8, how many scores fall: (a) above 76? (b) below 80? (c) above 50? (d) below 46? (e) between 48 and 52? (f) between 58 and 70?

9. Among normally distributed scores, what z score (a) divides the upper 5% of scores from the remainder? (b) divides the upper 2.5% of scores from the remainder? (c) divides the upper 1% of scores from the remainder? (d) divides the upper .5% of scores from the remainder?

10. Among normally distributed scores, what are the z score limits that identify the central (a) 99% of scores? (b) 95% of scores? (c) 50% of scores?

11. You have just learned that you scored 625 on the math portion of the Graduate Record Examination. The GRE has a mean of 500 and a standard deviation of 100. What percentage of students who took the GRE scored lower than you?

12. At Smart University, all new freshmen are given an English proficiency exam during the week of registration. Scores are normally distributed with a mean of 70 and a standard deviation of 10. The university has decided to place the top 25% into honors English and the bottom 20% into remedial English. What scores separate the upper 25% and lower 20% of the students from the rest?

13. General Machines Company is hiring new employees. The positions require some basic arithmetic skills, but the company is worried that the work is not challenging enough to keep highly skilled individuals interested. Therefore, they administer a screening test to all job applicants and decide to eliminate from consideration anyone scoring in the bottom or top 15%. If scores are normally distributed with a mean of 40 and a standard deviation of 8, what are the two outside limits between which applicants must score to be considered for the positions?

14. State University has no honors English course but instead places all freshmen scoring at least 60 on a proficiency exam (mean = 50, standard deviation = 10) into an advanced composition course. If the scores are normally distributed and there are 2,500 entering freshmen, how many will be put into the advanced course?

15. To join the national organization Mensa, you must provide verified evidence that you scored at least 150 on the Wechsler Adult Intelligence Scale (WAIS). WAIS scores are normally distributed in the population with a mean of 100 and a standard deviation of 15. (a) What proportion of the population is eligible to join? (b) If there are 180 million adult Americans, how many of them could join?

16. Seven thousand runners enter a local marathon in which they can qualify to enter the New York City Marathon if they complete the 26-mile-plus distance in under 3 hours and 10 minutes. Only 6,350 runners complete the race. If finishing times are normally distributed with a mean of 3 hours 40 minutes and a standard deviation of 28 minutes, how many runners qualified?

CHAPTER 7

Correlation

When you have finished studying this chapter, you should be able to:

- Understand that correlation refers to the degree of relationship between two variables, and the greater the correlation between two variables, the better we are able to predict the value of one from knowledge of the other;
- Construct a scatter diagram to graphically display the relationship between pairs of scores;
- Understand that the greater the degree of linear relationship between two variables, the more closely the points of intersection (for the X and Y variables) on the scatter diagram resemble a straight line;
- Understand that Pearson's product-moment correlation coefficient (r) is used to show only the extent of linear relationship between two variables;
- Explain how Pearson's r shows both the degree and direction of correlation between two variables, and know that values of r range from -1.00 (a perfect negative correlation) to $+1.00$ (a perfect positive correlation);
- Know how to calculate Pearson's r;

- Use IBM SPSS to calculate Pearson's *r*;
- Know when and how to use Spearman's rank-order correlation coefficient (r_S) in place of Pearson's *r*;
- Understand that a correlation between two variables does not prove that there is a causal relationship;
- Understand that adding or subtracting a (positive) constant to all scores, or multiplying or dividing all scores by a (positive) constant, does not affect the value of *r*; and
- Explain several variables that affect the value of *r*.

Remember the Saturday morning back in high school you spent hunched over the Scholastic Aptitude Test? (Maybe it was the American College Test.) The SAT is useful to colleges and universities because scores on the SAT (or the ACT) are related to grades earned later in college. Students with higher scores usually go on to do better in college than those with lower scores. However, the relationship is not perfect. Sometimes a person who scored low on the test gets surprisingly good grades in college; sometimes a person who scored high gets low grades.

How can we best describe the degree of relationship between two variables, such as test scores and college grades? This is an example of a question about **correlation**. Closely related is the problem of **prediction**. If there is a correlation between two variables, then from a person's score on one variable, we can do better than chance in predicting that person's score on the other variable. Many college admissions committees, for example, use SAT or ACT test scores to predict their applicants' freshman-year grade point averages (GPAs). If an applicant's predicted GPA falls below a certain standard, a committee may reject the applicant. As already indicated, the predictive value of the SAT is not perfect. In general, though, *the greater the association between two variables, the more accurately we can predict the standing in one from the standing in the other.*

correlation a measure of the degree of relationship between two variables

prediction when two variables are correlated, estimating the value of one variable from the known value of the other; the better the correlation between the variables, the better the prediction

Determining the degree of correlation and prediction is important in many areas in the behavioral sciences. To establish the reliability of a test (i.e., the consistency in scores over repeated administrations), for example, we would want to know to what extent initial performance correlates with performance on the same test at a later time. The test would be useless if it yielded scores that fluctuate widely over time. Correlation and prediction are also important in establishing test validity (i.e., the agreement between a test score and the quality it is supposed to measure). For example, to what extent are scores on a mechanical aptitude test predictive of on-the-job performance as a machinist? Other examples of problems involving association: Is the number of cigarettes smoked related to the incidence of lung cancer? Is the amount of exposure to violent themes in the media associated with aggressive behavior?

Correlation is the topic of this chapter, and prediction is the topic of the next. Chapter 9 will help enrich your understanding of both.

7.1 Some History

In 1859, Charles Darwin presented his theory of evolution by natural selection. One of the assumptions of the theory is that there is variation among individuals. To verify the theory, it became important to study traits in which organisms of the same species differ and to determine how those traits are influenced by heredity. These problems interested Darwin's cousin **Sir Francis Galton**, and Galton's research on them founded the study of individual differences.

To study the inheritance of stature, for example, Galton collected observations on families consisting of parents with grown children. To simplify his data, for each of the offspring, he calculated the midparent height as the mean of the mother's and the father's heights. (He first multiplied the mother's height by 1.08 to adjust it for the mean sex difference between the scores.) To organize and summarize the observations, Galton put the midparent height on one axis of a chart and offspring's height on the other axis in the manner displayed in Table 7.1. The data given in the table, in fact, are his (Galton, 1889, p. 208). Each entry in the body of the table is a frequency for that particular row–column intersection, called a *cell*. Look, for example, at the intersection of the row labeled 66–67 (height of offspring) and the column labeled 68–69 (midparent height). The frequency of 25 shown in that cell indicates that there were 25 children whose adult height was between 66 and 67 inches and whose midparent height was between 68 and 69 inches. Distributions like Table 7.1 that show the relation between two variables are called **bivariate distributions**.

bivariate distribution a distribution that shows the relation between two variables

From these data, Galton concluded that taller parents generally—but not inevitably—produced offspring who were also relatively tall. This relationship is apparent if we track the median offspring height column by column. The shaded cells are those in which the medians lie. For children whose midparent height was between 68 and 69 inches, for example, median adult height (found in the cell with 34 observations) was also between 68 and 69 inches. We can see that in the families with taller parents, represented in the columns toward the right of the table, the median offspring height was generally greater than in the families with shorter parents.

Galton perceived that he could summarize such data with a line showing how the children's heights related to the parents' heights, and that a straight line was often appropriate. We have drawn such a line in Table 7.1. Galton also saw that, other things being equal (such as sample variances and scale of measurement), a steeper line indicated a closer relationship between the two variables. In 1896, **Karl Pearson**, a scientist associated with Galton's laboratory, developed a rigorous mathematical treatment of these matters. Pearson's procedures are the ones you will learn about first.

7.2 Graphing Bivariate Distributions: The Scatter Diagram

The raw data summarized in Table 7.1 consisted of pairs of scores. Like Galton's study, *any problem in correlation requires pairs of scores*. For each case under observation, we must know the score on a first variable, which we will call X, and the score on a second variable, Y. The initial observations, therefore, can be listed in two columns. Table 7.2 offers an example. The subjects, designated by letters, are 10 male college students who filled out two questionnaires (unpublished data from Bear). One 10-item questionnaire measured the stress currently afflicting the student. Another 14-item questionnaire assessed difficulties in eating. We have used X for the total score on the stress questionnaire (higher scores indicate greater perceived stress) and Y for the total score on the eating difficulties questionnaire (higher scores indicate greater perceived difficulty).

To display bivariate data, modern researchers typically use a scatter diagram like Figure 7.1 (sometimes called a *scatterplot*), not a chart like Galton's (Table 7.1). Each point in a **scatter diagram** represents a single case and shows the scores on the two variables for that case. For example, in Figure 7.1, which plots the X and Y scores from Table 7.2, the dot in the upper right corner represents student J, who scored 30 on the stress test (variable X) and 28 on the eating difficulties test (variable Y). The steps in constructing a scatter diagram from a set of paired scores are summarized as follows:

scatter diagram a graph of a bivariate distribution consisting of dots at the point of intersection of paired scores

Step 1: Designate one variable X and the other Y. Although it does not matter which is which, in cases where one variable is used to predict the other, we will follow the custom of using X for the "predictor" variable.

LibraryPirate

TABLE 7.1 *Galton's Data on Midparent Height and Height of Adult Offspring*

HEIGHT OF ADULT OFFSPRING* (IN.)	MIDPARENT HEIGHT (IN.)										
	BELOW 64	64–65	65–66	66–67	67–68	68–69	69–70	70–71	71–72	72–73	ABOVE 73
Above 74							5	3	2	4	
73–74						3	4	3	2	2	3
72–73			1		4	4	11	4	9	7	
71–72			2		11	18	20	7	4	2	
70–71			5	4	19	21	25	14	10	1	
69–70	1	2	7	13	38	48	33	18	5	2	
68–69	1		7	14	28	34	20	12	3	1	
67–68	2	5	11	17	38	31	27	3	4		
66–67	2	5	11	17	36	25	17	1	3		
65–66	1	1	7	2	15	16	4	1	1		
64–65	4	4	5	5	14	11	16				
63–64	2	4	9	3	5	7	1	1			
62–63		1		3	3						
Below 62	1	1	1			1		1			

*All female heights multiplied by 1.08 to adjust for sex differences.

TABLE 7.2 *Bivariate Data: Scores for 10 Male College Students on Two Self-Report Measures*

STUDENT	STRESS *X*	EATING DIFFICULTIES *Y*
A	17	9
B	8	13
C	8	7
D	20	18
E	14	11
F	7	2
G	21	5
H	22	15
I	19	26
J	30	28

Step 2: Draw axes of about equal length for your graph. Note that this differs from the convention used for constructing frequency histograms or polygons.

Step 3: Place the high values of *X* to the right on the horizontal axis and the high values of *Y* toward the top on the vertical axis. Label convenient points along each axis.

Step 4: For each pair of scores, find the point of intersection for the *X* and *Y* values and indicate it with a dot. Where dots exactly overlap, substitute for them the integer telling how many cases occur in that spot (as Galton did in Table 7.1).

Step 5: Name each axis and give the entire graph a name.

linear relationship a relationship that can best be represented by a straight line

The scatter diagram allows us to easily see the nature of the relationship, if any exists, between the two variables. Pearson's mathematical procedures assume an underlying **linear relationship**, that is, a relationship that can be best represented by a straight line. Figure 7.1

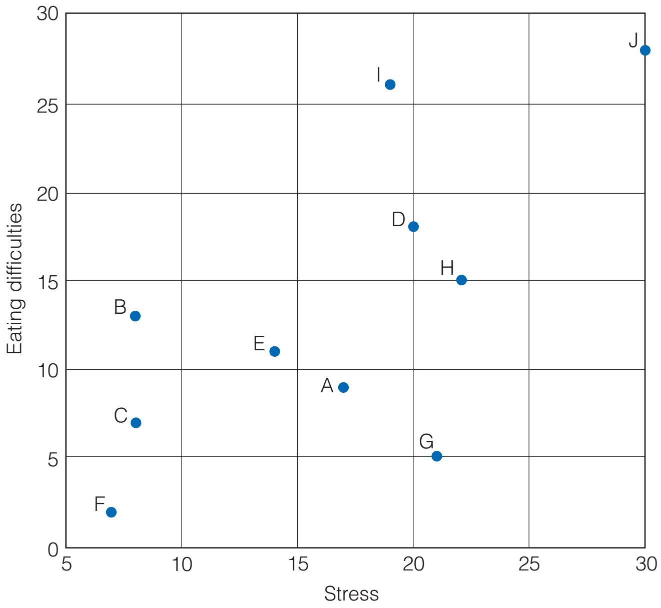


FIGURE 7.1 Scatter diagram of bivariate distribution of stress scores and eating difficulties scores for 10 male college students. Data from Table 7.2.

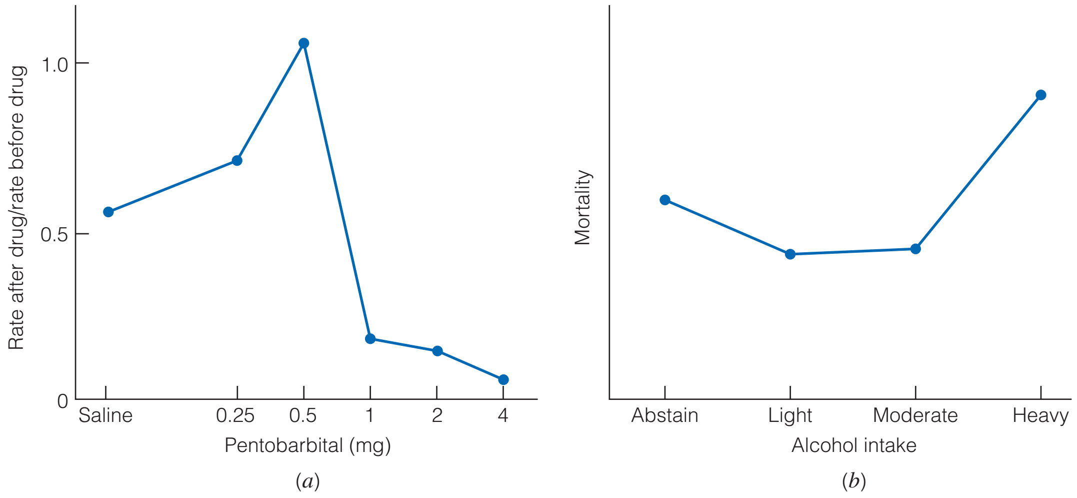


FIGURE 7.2 Examples of curvilinear relationships. (*a*) Effects of pentobarbital on lever-pressing for food reinforcement in pigeons (modified from Dews, 1955). (*b*) Relationship between weekly alcohol intake and mortality (modified from San José et al., 1999).

shows such a relationship. As a general rule, those students who scored high on the stress questionnaire also had high scores on the eating difficulties questionnaire. The relationship is not perfect. Person G, for example, had a total score of 21 (the third highest) on the stress questionnaire, but had a total score of only 5 (the second lowest) on the eating difficulties questionnaire. Nevertheless, a quick inspection of the graph shows us that Pearson's assumption of linearity is appropriate here.

It must be emphasized that not all relationships are linear. Figure 7.2*a*, for example, displays the effects of dosage levels of pentobarbital, a barbiturate, on lever-pressing for food reinforcement in pigeons. Figure 7.2*b* shows the relationship between weekly alcohol consumption and mortality. Because the relationships here are **curvilinear** (the points on a scatter diagram cluster about a curved line), Pearson's procedures are not appropriate (and, in fact, will give misleading results). *Except where noted, the techniques of correlation and prediction presented in this text are appropriate only where the underlying relationship between two variables is linear.*

7.3 Correlation: A Matter of Direction

positive correlation a linear relationship in which high scores on the first variable are generally paired with high scores on the second, and low scores on the first variable are generally paired with low scores on the second

The scatter diagram not only shows whether there is a linear relationship between two variables but also shows at a glance the direction of the relationship. If the higher scores on X are generally paired with the higher scores on Y, and the lower scores on X are generally paired with the lower ones on Y, then the direction of the correlation between the two variables is positive. In a scatter diagram, a **positive correlation** appears as a cluster of data points that slopes from lower left to upper right. Thus, the correlation shown in Figure 7.1 is positive. It would not matter if we had designated the stress scores as Y instead of X (and plotted them along the vertical axis instead of the horizontal) and the eating difficulties scores as X instead of Y (and plotted them along the horizontal axis): the dots in the scatter diagram would still slope from lower left to upper right. Try it. Figure 7.3 offers more examples of positive correlations.

If the higher scores on X are generally paired with the lower scores on Y and the lower scores on X are generally paired with the higher ones on Y, then the direction of the correlation is negative. For example, a study conducted by the California Department of Education

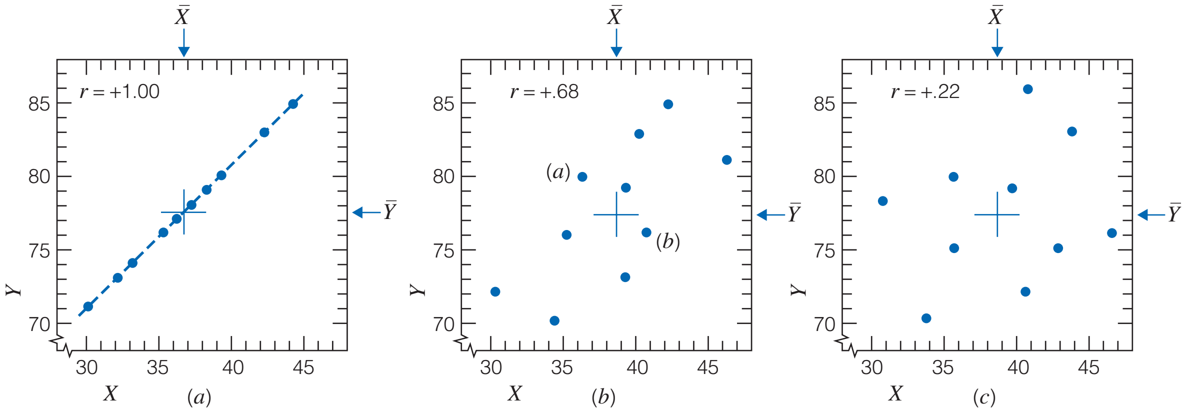


FIGURE 7.3 Scatter diagrams illustrating various degrees of positive (+) correlation.

found a negative correlation between hours of TV watched daily by children and their scores on standardized reading and mathematics tests. As the number of hours spent watching television increased, test scores decreased. In a scatter diagram, a **negative correlation** appears as a cluster of data points that slopes from upper left to lower right, as illustrated in Figure 7.4*b* and 7.4*c*. Again, it does not matter which variable is plotted along the horizontal axis and which along the vertical axis; the direction of the slope will be the same.

In cases where there is **no correlation** between two variables (i.e., both high and low values of *X* are equally paired with both high and low values of *Y*), there is no direction in the pattern of the dots. They are scattered about the diagram in an irregular fashion, as illustrated in Figure 7.4*a*.

negative correlation
a linear relationship in which high scores on the first variable are generally paired with low scores on the second, and vice versa

7.4 Correlation: A Matter of Degree

Look closely again at Figure 7.3. All three scatter diagrams show a positive correlation between *X* and *Y*. You can tell because the dots tend to go from lower left to upper right. Then how do these correlations differ? They differ in degree. The scatter diagram in Figure 7.3*a* shows a perfect linear correlation, one where all the data points fall on a straight line. The other correlations are less than perfect, but in each case one can still construct (as Galton did) through the cluster of dots a straight line that summarizes the relationship between the two variables. With less than perfect correlations, the dots will show some scatter about the straight line. The more scatter, the weaker the correlation.

The degree of association shared by two variables is indicated by the **coefficient of correlation**, invented by Pearson in 1896. Galton was the first to use the symbol ***r*** for a *sample correlation coefficient*; it is often written with subscripts as r_{XY}. It is calculated in a way we will show later, from the *pairs* of scores. When a perfect correlation exists, its value is plus or minus 1.0. When no linear relationship exists, its value is 0. Thus, intermediate degrees of correlation are represented by values between 0 and ±1.0.

The sign (plus or minus) indicates only the direction of the correlation (+ for a positive correlation and − for a negative correlation), not its degree. Students sometimes mistakenly believe that a correlation coefficient with a positive value indicates a stronger degree of relationship than does a coefficient with a negative value, but this is not so. Two correlation coefficients having the same absolute value but differing in sign indicate the same degree of relationship. Only the direction of the relationship differs. Thus, a correlation

no correlation
both high and low values on the first variable are equally paired with both high and low values on the second variable

coefficient of correlation
a mathematical expression of the degree of linear association between two variables

r
symbol for Pearson's coefficient of correlation

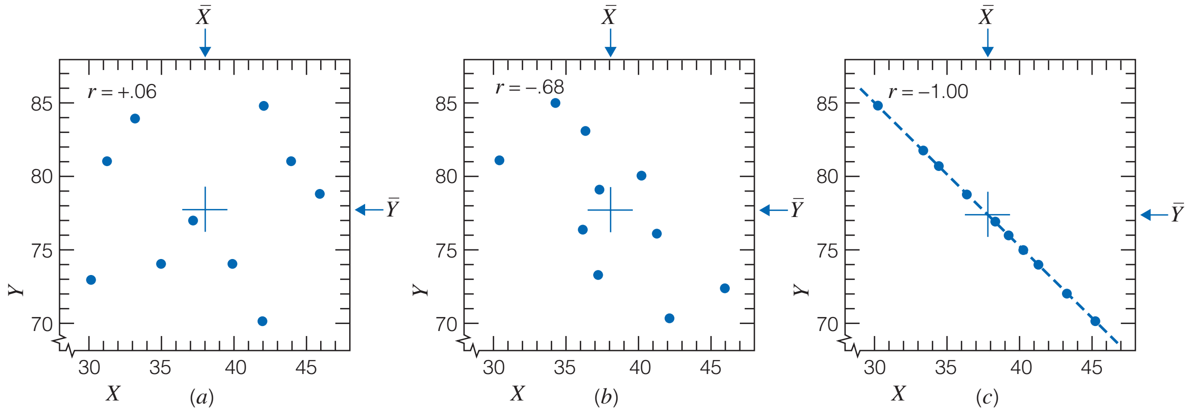


FIGURE 7.4 Scatter diagrams illustrating (*a*) an essentially zero correlation and (*b* and *c*) two degrees of negative (−) correlation.

coefficient of −.50 indicates just as strong a relationship between two variables as a coefficient of +.50.

To summarize the relationship between a scatter diagram and the correlation coefficient, *Pearson's correlation coefficient is a number that indicates how well the data points in a scatter diagram "hug" the straight line of best fit.* With perfect correlations, all the data points fall exactly on a straight line as in Figures 7.3*a* and 7.4*c* and the value of the coefficient is ±1.0. When the association between two variables is less than perfect, the data points show some scatter about the straight line that summarizes the relationship, and the *absolute* value of the correlation coefficient is less than 1.0. The weaker the relationship, the more scatter and the lower the absolute value of the correlation coefficient.

In the real world, perfect correlations occur only in trivial instances; for example, the correlation will be −1.00 between the number of correct answers on a test and the number of errors plus omissions. Table 7.3 lists typical **values of *r*** for some correlations in the behavioral sciences. We began this chapter by noting that there is a relationship between SAT scores and freshman-year GPAs. In fact, SAT scores typically correlate about +.35 with freshman-year GPAs. The value of the coefficient indicates a modest degree of relationship.

values of *r* range from −1.00 (a perfect negative correlation) to 0.0 (no correlation) to +1.00 (a perfect positive correlation)

TABLE 7.3 *Typical Values of* r

VARIABLES	*r*
IQ from one form of Wechsler Adult Intelligence Scale (WAIS) and IQ from an alternate form	+.90
Childhood IQ and adult IQ	+.70 to +.85
Father's years of education and grown child's years of education	+.60
Verbal score on Scholastic Aptitude Test (SAT) and mathematics score on SAT	+.60
IQ of husband and IQ of wife	+.50
Total score on SAT and freshman-year grade point average (GPA)	+.35
Total score on Graduate Record Exam (GRE) and undergraduate GPA among applicants to a highly selective graduate program in psychology	+.20
Attitudes about school and cutting school among junior high and high school students	−.29
Latency of visual evoked response and conceptual age at time of birth	−.61

Sources: Information on WAIS from Matarazzo (1972); on childhood IQ from McCall (1977); on SAT and GPA from Slack and Porter (1980); on GRE and GPA from Dawes (1975); on attitudes about school from Epstein and McPartland (1976); on visual evoked response and conceptual age from Engel and Benson (1968).

7.5 Understanding the Meaning of Degree of Correlation

Although the absolute value of r does represent the degree of linear relationship between pairs of scores, the meaning is complex. A correlation between X and Y does not mean that variation in X causes variation in Y, or vice versa (more about this in Section 7.10). Moreover, a correlation coefficient of .50 *does not* even indicate "50% association." *Correlations are not percentages*. Nor does a correlation of .50 indicate twice the strength of association as a correlation of .25. What does it mean, then, to say that there is a +.35 correlation between SAT scores and freshman-year GPAs?

Many students with low or mediocre SAT scores who are doing well in school may decide that what they learn in statistics class has little to do with the real world. Whenever there is less than perfect correlation between two variables, there will be exceptions. One way to appreciate the meaning of size differences in correlation coefficients is to understand how often exceptions are expected with different degrees of correlation. We can do this by comparing the percentage of cases scoring above the median on the first variable with the percentage scoring above the median on the second variable. With perfect correlations, all the cases above the median for the first variable are also above the median for the second. When there is zero correlation (no association), only 50% of the cases above the median on the first variable will also be above the median on the second. This indicates, of course, that half of those above the median on the first variable will be below the median on the second.

Table 7.4 shows us what we can expect with different size correlations. For the correlation of +.35 between SAT scores and freshman-year GPAs, the table tells us that of those who scored above the median on the SAT, we can expect that 61.2% will also be above the median for GPA and that 38.8% will have GPAs lower than the median. This is better than chance (50–50), but far from perfect. The data in the table assume that the variables are continuous, normally distributed, and describe an entire population. However, even

TABLE 7.4 *The Meaning of True Correlations*[a]

TRUE CORRELATION	CASES EXPECTED ON SECOND VARIABLE (%)	
	Above median	*Below median*
0.00	50.0	50.0
0.10	53.1	46.9
0.20	56.2	43.8
0.30	59.5	40.5
0.35	**61.2**	**38.8**
0.40	63.0	37.0
0.50	66.5	33.5
0.60	70.3	29.7
0.70	74.5	25.5
0.80	79.3	20.7
0.90	85.3	14.7
1.00	100.0	0.0

[a] For cases above the median on variable 1, this table gives the expected percentages of cases falling above and below the median on variable 2 for true correlations. Column entries are reversed for negative correlations.

Source: Credit is owed to W. B. Michael, "An Interpretation of the Coefficients of Predictive Validity and of Determination in Terms of the Proportions of Correct Inclusions or Exclusions in Cells of a Fourfold Table," *Educational and Psychological Measurement*, **26**, no. 2 (1966): 419–24.

with modestly sized samples, the figures provide ballpark estimates. As a first step this table should provide some help in understanding the meaning of degree of correlation. We provide a more technical explanation of what correlations measure in Chapter 9 (see Sections 9.2–9.3 and 9.4–9.6).

7.6 Formulas for Pearson's Coefficient of Correlation

In this section, we introduce formulas for computing Pearson's coefficient of correlation. There are other indices of association suited for special situations, but Pearson's is by far the most common. In fact, when researchers speak of a correlation coefficient without being specific about which one they mean, you may safely assume they are referring to Pearson's r. Technically, it is a **product-moment correlation coefficient**.[1]

A correlation, you recall, shows the relationship between pairs of scores. In its simplest form, the Pearson correlation coefficient is defined as follows:

$$\boldsymbol{r = \frac{\Sigma(z_X z_Y)}{n}}$$

z-score formula for Pearson's r

z-SCORE FORMULA FOR PEARSON'S CORRELATION COEFFICIENT

$$r = \frac{\sum(z_X z_Y)}{n} \tag{7.1}$$

where n is the number of pairs of scores. To find r, we must convert each raw score to a z score. Then for each pair of z scores, we multiply the scores, sum the results, called the **cross-products**, and divide by the number of pairs of scores. Thus, r is the mean of the cross-products of the paired z scores.

We can also define Pearson's coefficient directly in terms of deviation scores:

$$\boldsymbol{r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{nS_X S_Y}}$$

deviation-score formula for Pearson's r

DEVIATION-SCORE FORMULA FOR PEARSON'S CORRELATION COEFFICIENT

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{nS_X S_Y} \tag{7.2}$$

where n is again the number of pairs of scores and S_X and S_Y are the standard deviations of the two samples. (Note 7.1 shows how Formula 7.2 is derived from Formula 7.1.)

n
in formulas for a correlation coefficient, the number of pairs of scores

When calculating r by hand, you would probably never use Formulas 7.1 or 7.2 (because it is too much work). We introduce them here only because it may help you to better understand Pearson's coefficient. For example, look carefully at the numerator for Formula 7.2. Because the denominator always has a positive value (n, S_X, and S_Y are always positive numbers), it is apparent that the sum of the cross-products (of the paired deviation scores) determines whether the coefficient will be negative, zero, or positive. It will be zero when the sum of the negative cross-products equals the sum of the positive cross-products. The coefficient will be negative when the sum of the negative cross-products exceeds the sum of the positive cross-products, and it will be positive when the reverse is true. Typically, the greater the predominance of cross-products bearing one sign over those bearing the other, the greater the magnitude of the coefficient.

---

[1] The term *moment* is borrowed from physics and refers to a function of the distance of an object from the center of gravity. With a frequency distribution, the mean is the center of gravity and, thus, deviation scores are the moments. As you will see in Formula 7.2, Pearson's r is calculated by taking the *products of the paired moments.*

The magnitude of the coefficient really depends on the similarity of the paired z scores. Look again at Formula 7.1. The sum of the cross-products ($\Sigma z_X z_Y$) is maximal (equal to Σz_X^2 or Σz_Y^2) only when both scores in each pair are at the same relative distance from their means (that is, when the magnitudes of the two z scores are the same). When both scores in each pair do share the same relative position, the result is a straight-line relationship and a maximum value for the correlation coefficient. Thus, the correlation coefficient is really an index of how similar the paired z scores are.

7.7 Calculating *r* from Raw Scores

The z-score formula and the deviation-score formula for r are useful for teaching purposes, but they are time consuming if you have to calculate r by hand. However, it is easy to transform the deviation-score formula to a useful computing formula:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{nS_XS_Y} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\cancel{n}\sqrt{\dfrac{\sum (X - \overline{X})^2}{\cancel{n}}}\sqrt{\dfrac{\sum (Y - \overline{Y})^2}{\cancel{n}}}}$$

$$= \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 (Y - \overline{Y})^2}}$$

Thus,

COMPUTATIONAL FORMULA FOR PEARSON'S
CORRELATION COEFFICIENT

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{(SS_X)(SS_Y)}} \qquad (7.3)$$

The two expressions in the denominator are, of course, the sums of squares of the deviation scores for X and Y. In Section 5.5 you learned how to compute these quantities from raw scores. Here are the formulas again:

$$SS_X = \sum (X - \overline{X})^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

and

$$SS_Y = \sum (Y - \overline{Y})^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

The numerator of Formula 7.3, $\Sigma(X - \overline{X})(Y - \overline{Y})$, is the sum of cross-products of deviation scores and can also be expressed in terms of raw scores.

TABLE 7.5 *Calculation of* r *from the Raw Scores of Table 7.2*

STUDENT	X	Y	X^2	Y^2	XY
A	17	9	289	81	153
B	8	13	64	169	104
C	8	7	64	49	56
D	20	18	400	324	360
E	14	11	196	121	154
F	7	2	49	4	14
G	21	5	441	25	105
H	22	15	484	225	330
I	19	26	361	676	494
J	30	28	900	784	840
$n = 10$	Sum: 166	134	3,248	2,458	2,610

①

② $SS_X = \sum X^2 - \left(\sum X\right)^2/n = 3{,}248 - 166^2/10 = 492.4$

③ $SS_Y = \sum Y^2 - \left(\sum Y\right)^2/n = 2{,}458 - 134^2/10 = 662.4$

④ $\sum (X - \overline{X})(Y - \overline{Y}) = \sum XY - \dfrac{\left(\sum X\right)\left(\sum Y\right)}{n} = 2{,}610 - \dfrac{(166)(134)}{10} = 385.6$

⑤ $r = \dfrac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{(SS_X)(SS_Y)}} = \dfrac{385.6}{\sqrt{(492.4)(662.4)}} = \dfrac{385.6}{\sqrt{326{,}165.76}} = \dfrac{385.6}{571.11} = +.675$

RAW-SCORE EQUIVALENT
OF $\Sigma(X - \overline{X})(Y - \overline{Y})$

$$\sum (X - \overline{X})(Y - \overline{Y}) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} \qquad (7.4)$$

We show how this formula is derived in Note 7.2 at the end of the chapter.

We illustrate the raw-score method of calculating Pearson's r in Table 7.5, using the data from Table 7.2 again. The steps are summarized as follows:

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{(SS_X)(SS_Y)}}$$

computational raw-score formula for Pearson's r

Step 1: Find the six building blocks from which everything else is generated: n, ΣX, ΣY, ΣX^2, ΣY^2, ΣXY. The outcomes are indicated by ① in Table 7.5.

Step 2: Find SS_X, SS_Y, and $\Sigma(X - \overline{X})(Y - \overline{Y})$. These are shown at ②, ③, and ④.

Step 3: Substitute these last three values in Formula 7.3 and do the remaining arithmetic ⑤.

The positive correlation we have obtained tells us that it was at least generally true that those with higher perceived stress also had greater perceived eating difficulties. As a check on your work, always look at the scatter diagram after you have calculated r. Remember that with a positive correlation, the cluster of dots should go from the lower left to the upper right. If they do not, you have made a mistake in your calculations. In this example, the value

+.675 indicates that the relationship between the two variables was a fairly close one, but certainly not perfect. So, low perceived stress did not necessarily guarantee a lack of eating difficulties, just as high stress did not guarantee that the student had eating difficulties (witness Student G).

7.8 Calculating *r* with IBM SPSS

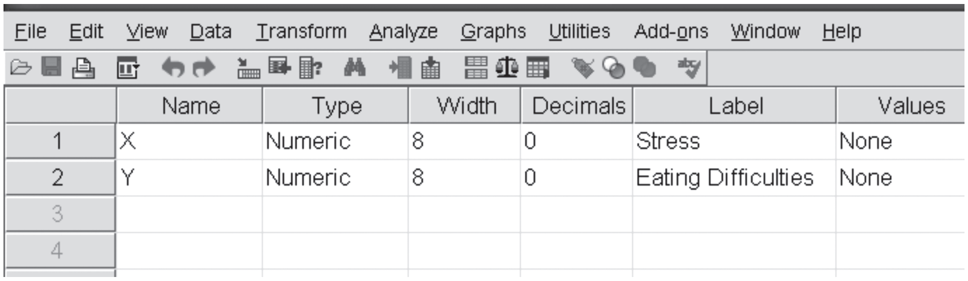
In Section 5.6, you learned how to obtain the standard deviation of a variable in IBM SPSS. Now you will learn how to use IBM SPSS to compute Pearson's *r*.

Recall that after you open IBM SPSS, a dialog box will appear with the following question: "What would you like to do?" Because you will be entering data, select "Type in data" and click the OK button. For future reference, if you have a IBM SPSS data file that you would like to open, here is how you can access it. Specifically, after opening IBM SPSS, instead of selecting "Type in data," you should select "Open an existing data source." Then browse for your data file and select it as the data source to be opened.

After clicking the OK button, click the **Variable View** tab located toward the bottom left of the screen. Then, in the first column labeled Name, type the variable names—X in the first row and Y in the second. For numerical variables, you can use the fourth column in **Variable View** labeled Decimals to change the number of decimals displayed in the **Data Editor**. Note that changing the number of decimals displayed does not change the actual values that you enter in the **Data Editor**. Because the data for X and Y in Table 7.2 are all whole numbers, there is little reason to display the 0s to the right of the decimal place. So, change the number of decimals displayed from 2 to 0. In the next column, you can type a descriptive label for your variables. You should type Stress in the first row and Eating Difficulties in the second. Figure 7.5 shows a screenshot of the **Variable View** after typing the variable names, changing the number of decimals displayed to 0, and entering descriptive labels. Now, click the **Data Editor** tab and enter the paired scores shown in Table 7.2.

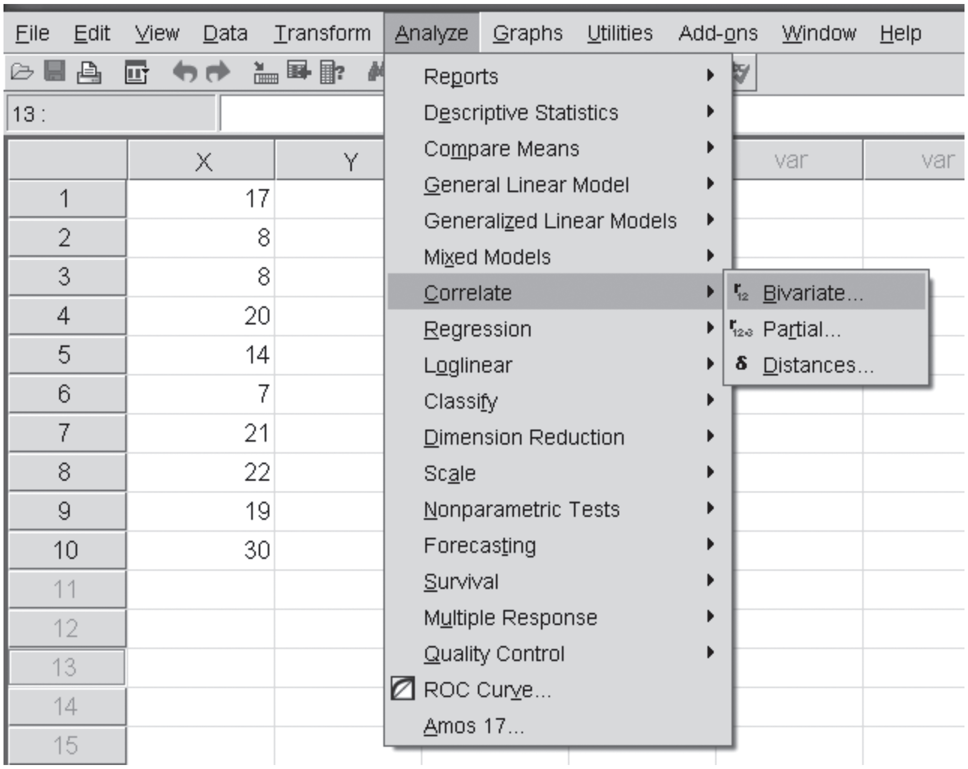
To obtain correlation coefficients in IBM SPSS, click **Analyze** and then select **Correlate** followed by **Bivariate**. Figure 7.6 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze** → **Correlate** → **Bivariate**.

Now a dialog box will appear like that shown in Figure 7.7. Notice that the descriptive labels and their variable names appear in the box on the left. You should move the variables you want to correlate by highlighting them and clicking on the arrow button that appears in the center of the dialog box. Figure 7.8 shows a screenshot after the X and Y variables have been moved to the box on the right. You may have noticed in the figure that Pearson's is the



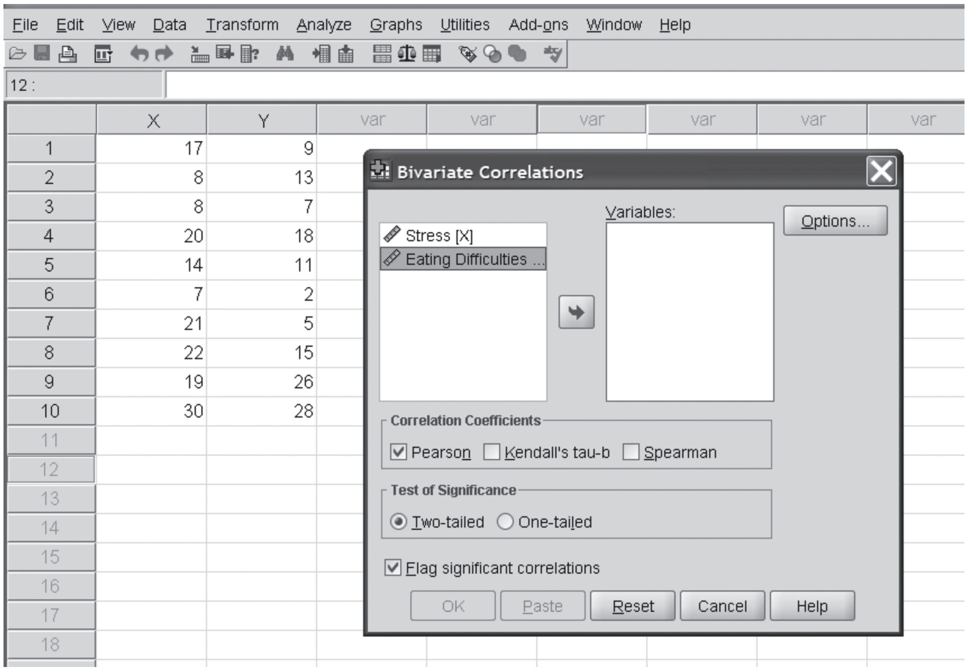
	Name	Type	Width	Decimals	Label	Values
1	X	Numeric	8	0	Stress	None
2	Y	Numeric	8	0	Eating Difficulties	None
3						
4						

FIGURE 7.5 Variable view in IBM SPSS.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
13 :
var
var
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Bivariate...
Partial...
Distances...

	X	Y
1	17	
2	8	
3	8	
4	20	
5	14	
6	7	
7	21	
8	22	
9	19	
10	30	
11		
12		
13		
14		
15		

FIGURE 7.6 IBM SPSS menus when calculating a correlation.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
12 :
var
var
var
var
var
var
Bivariate Correlations
Variables:
Options...
Stress [X]
Eating Difficulties ...
Correlation Coefficients
Test of Significance
OK
Paste
Reset
Cancel
Help

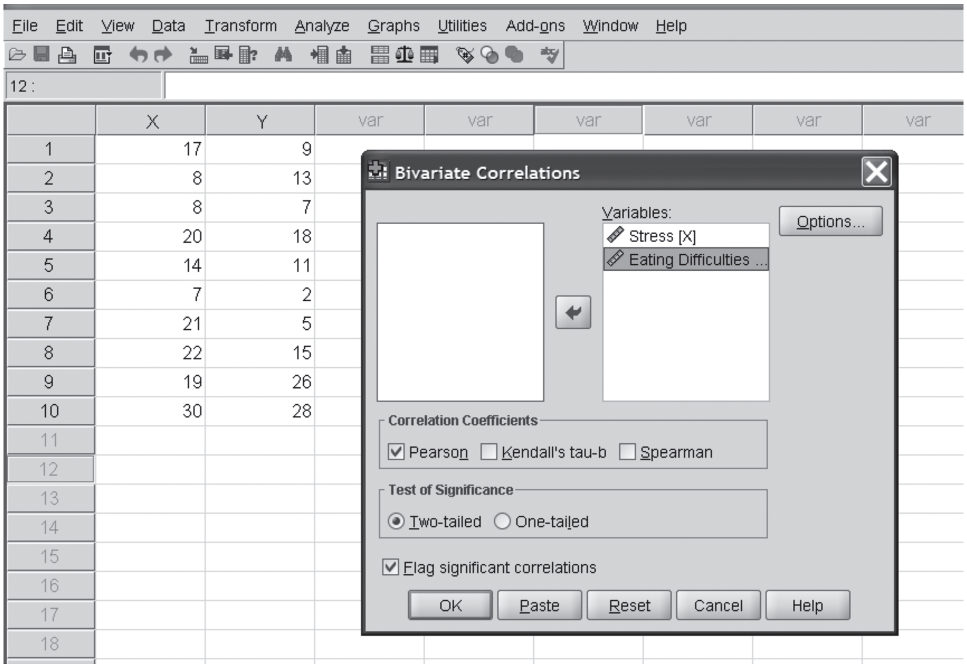
	X	Y
1	17	9
2	8	13
3	8	7
4	20	18
5	14	11
6	7	2
7	21	5
8	22	15
9	19	26
10	30	28
11		
12		
13		
14		
15		
16		
17		
18		

Correlation Coefficients: ☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

FIGURE 7.7 IBM SPSS dialog box when calculating a correlation.



	X	Y
1	17	9
2	8	13
3	8	7
4	20	18
5	14	11
6	7	2
7	21	5
8	22	15
9	19	26
10	30	28

Bivariate Correlations

Variables:
Stress [X]
Eating Difficulties ...

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

FIGURE 7.8 IBM SPSS dialog box when calculating a correlation after selecting variables.

type of correlation coefficient that IBM SPSS computes by default. Thus, we will keep the default settings. Now, click the OK button. Table 7.6 shows the IBM SPSS output and confirms that Pearson's *r* is +.675.

Notice that Table 7.6 shows "Sig. (2-tailed)." This has to do with hypothesis testing, which we will introduce in Chapter 12. You will learn about testing hypotheses for *r* in Chapter 16.

TABLE 7.6 *IBM SPSS Output when Calculating a Correlation*

```
CORRELATIONS
  /VARIABLES = X Y
  /PRINT = TWOTAIL NOSIG
  /MISSING = PAIRWISE.
```

Correlations

		Stress	Eating Difficulties
Stress	Pearson Correlation	1	.675*
	Sig. (2-tailed)		.032
	N	10	10
Eating Difficulties	Pearson Correlation	.675*	1
	Sig. (2-tailed)	.032	
	N	10	10

*Correlation is significant at the 0.05 level (2-tailed).

7.9 Spearman's Rank-Order Correlation Coefficient

r_S
symbol for Spearman's rank-order correlation coefficient

Recall from Section 1.6 that in the behavioral sciences there are many measuring instruments that lack equal intervals and an absolute zero point. We can, however, place the observations in rank order (see the ordinal scale, Section 1.5). **Spearman's rank-order correlation coefficient**, **r_S**, is closely related to the Pearson correlation coefficient.[2] In fact, if the paired scores are both in the form of ranks (and there are no ties in rank), calculation of r_S and Pearson's r result in identical outcomes.

When would one want to use the rank-order coefficient? Suppose 10 aspects of a job are identified, such as hours of work, working conditions, quality of supervision, wages, and so forth. Suppose we ask workers to place these job aspects in rank order according to their importance, and suppose we also ask their supervisors to rank the same aspects according to the importance that they believe workers would assign. We may now ask about the extent to which one worker agrees with another and the extent to which supervisors understand workers' feelings. We can study the degree of agreement between any two persons by calculating the coefficient of correlation between the ranks they assign to the 10 job aspects. Because the data are in the form of ranks, r and r_S will yield the same coefficient. We prefer r_S in this case because it is simpler to calculate.

We can also use r_S when measures are in score form. In this case, we translate each set of measures into rank form, assigning 1 to the lowest score, 2 to the next lowest, and so on. When would we do this? Sometimes the scale properties of the measures appear doubtful (see Section 1.6). If what matters is that one score is higher than another and how much higher is not really important, translating scores to ranks will be suitable. In any event, we typically use r_S only in circumstances in which n is rather small. When n is large, the proportion of tied ranks is likely to increase, and the work of translating scores to ranks becomes progressively more burdensome and error-prone.

The formula for the Spearman rank-order correlation coefficient is:

$r_S = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$
formula for Spearman's r_S

SPEARMAN RANK-ORDER
CORRELATION COEFFICIENT

$$r_S = 1 - \frac{6\sum D^2}{n(n^2 - 1)} \qquad (7.5)$$

where: D = the difference between a pair of ranks
n = the number of pairs of ranks

Before we proceed, you must first learn how to place scores in rank order. Most rank-order procedures in statistics require that the underlying variable be continuous. However, many dependent measures in the behavioral sciences are discrete variables (such as scores on a multiple-choice test), which often results in identical scores. This poses a problem because identical scores lead to ties in rank.

There are various ways to deal with ties in rank. A simple and reasonably satisfactory way is to assign to each of the tied scores the mean of the ranks that would be available to them. Consider the following nine scores:

4 5 5 8 11 11 11 15 19

---

[2] The Spearman rank-order correlation coefficient is sometimes called Spearman's rho and symbolized ρ. In this book, ρ is reserved for the population value of Pearson's r.

By this principle, we would assign these ranks:

1 2.5 2.5 4 6 6 6 8 9

This procedure usually has little or no effect on the mean, but it tends to reduce the variance. The effect of this on the statistic being calculated is usually slight *unless* a third or more of all scores are involved in ties.

When scores are fairly numerous, and especially when there are ties, it is easy to make a mistake in assigning ranks. We therefore use the following check:

$$\textit{sum of the ranks} = \frac{n(n+1)}{2}$$

where: $n =$ the number of scores being ranked.

This check holds whether or not there are ties.

Now let's try an example problem. Suppose an instructor is curious about the relation between the order in which the 15 members of her class completed an examination and the number of points earned on it. After the test, the examination booklets form a pile on her desk; the bottom booklet is the first one turned in and the remainder are in the order in which they were submitted. She assigns a rank of 1 to the first paper turned in and succeeding ranks according to the order of completion. After she has scored the tests, she records the order of turn-in (①) and the score obtained (②), as shown in Table 7.7. She then converts the test scores to ranks, assigning a rank of 1 to the lowest score. Because two scores are tied, the instructor assigns the average of the ranks available for them to each. The set of paired ranks appears in the columns headed R_X (③) and R_Y (④). Next, she records the differences between the pairs of ranks in the column headed D (⑤). If the algebraic signs are retained, a useful check may be based on the fact that ΣD must be zero. The instructor then squares and sums the values of D (⑥) and completes the calculation of r_S as shown in the table (⑦). The value of r_S is $+.38$.

TABLE 7.7 *Calculation of* r_S

SUBJECT	① ORDER OF TURN-IN X	② TEST SCORE Y	③ RANK OF X R_X	④ RANK OF Y R_Y	⑤ $D = R_X - R_Y$	⑥ D^2
A	1	28	1	6.5	−5.5	30.25
B	2	21	2	2	0.0	0.00
C	3	22	3	3.5	−.5	.25
D	4	22	4	3.5	.5	.25
E	5	32	5	10	−5.0	25.00
F	6	36	6	13	−7.0	49.00
G	7	33	7	11	−4.0	16.00
H	8	39	8	15	−7.0	49.00
I	9	25	9	5	4.0	16.00
J	10	30	10	8	2.0	4.00
K	11	20	11	1	10.0	100.00
L	12	28	12	6.5	5.5	30.25
M	13	31	13	9	4.0	16.00
N	14	38	14	14	0.0	0.00
O	15	34	15	12	3.0	9.00
$n = 15$						$\sum D^2 = 345.00$

Calculation: ⑦ $r_S = 1 - \dfrac{6\sum D^2}{n(n^2-1)} = 1 - \dfrac{6(345)}{15(15^2-1)} = .38$

7.10 Correlation Does Not Prove Causation

If variation in *X* causes variation in *Y*, that causal connection will appear in some degree of correlation between *X* and *Y* (at least when we remove any obscuring complications). However, we cannot reason backward from a correlation to a causal relationship. The fact that *X* and *Y* vary together is a necessary but not a sufficient condition for us to conclude that there is a cause-and-effect connection between them.

To appreciate this point, we will review a study by psychologist Sandra Scarr (1985). Scarr wanted to know whether the technique used by a mother to discipline her children was related to the children's later social adjustment, intelligence, and communication skills. On the basis of interviews with and observations of the mothers, she distinguished several measures of discipline ranging from negative (including things like physical punishment) to very positive (including reasoning and explaining things to the child). Good disciplinary technique was positively correlated with all three of the tested variables. Children who had been raised by mothers using good discipline techniques were better adjusted, more intelligent, and more communicative. The "obvious" conclusion is that the good discipline techniques caused the children's higher IQs and better social skills.

Do the positive correlations found by Scarr really provide evidence of cause-and-effect relationships? No, for there may have been other variables affecting both the mothers' use of discipline and the children's scores. In fact, when Scarr measured the mothers' IQs, she found that they correlated positively with both. After using other statistical techniques to remove the factor of the mothers' IQs, Scarr no longer found a positive correlation between their discipline techniques and their children's later IQs and behavioral skills. Mothers with higher IQs simply tended to prefer positive discipline techniques and to have children with higher IQs and better skills.

Here is another example: A 1992 study of 11,000 adults found a positive correlation between vitamin C intake and life expectancy. Men with a daily intake of 150 mg vitamin C lived 2 years longer than men in the lowest intake group, whereas men with a daily intake of 300 mg lived 6 years longer. The men who took vitamin C had much lower rates of heart disease. Can we conclude from this that vitamin C causes lower rates of heart disease and increased life span? Many doctors say probably not, because men who take lots of vitamin C are also likely to take better care of themselves in other ways as well—exercise more, smoke less, eat more vegetables and fruit, and so on.

Figure 7.9 shows four of the possibilities that may underlie a correlation between two variables: (*a*) The condition of *X* determines (in part, at least) the condition of *Y*. (*b*) The opposite is true—*Y* is a cause of *X*. (*c*) Some third variable influences both *X* and *Y*. (*d*) A complex of interrelated variables influences *X* and *Y*. Moreover, two or more of these situations may occur simultaneously. For example, *X* and *Y* may influence each other. Thus, an investigator must be alert to the several possible interpretations of a correlation.

Although a correlation by itself does not prove causation, we do not wish to leave you with the impression that correlations are of no importance in research. The finding of a strong

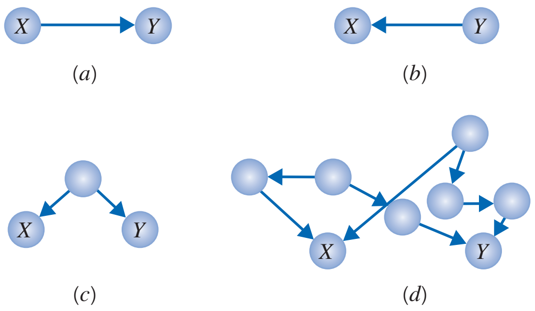
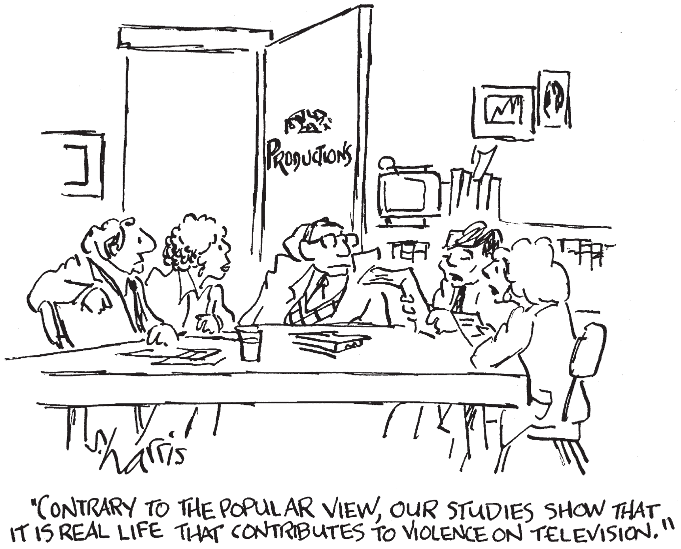


FIGURE 7.9 Possible relationships between variables *X* and *Y* that may underlie a correlation.



"CONTRARY TO THE POPULAR VIEW, OUR STUDIES SHOW THAT IT IS REAL LIFE THAT CONTRIBUTES TO VIOLENCE ON TELEVISION."

© Sidney Harris, ScienceCartoonsPlus.com

correlation between two variables is often the starting point for additional research. In 1955, for example, a researcher found a correlation of +.70 between the mean number of cigarettes smoked and the incidence of lung cancer in 11 countries (Doll, 1955). The correlation alone did not prove causation, for the countries also differed in levels of air pollution and other factors that might cause lung cancer, but the results were nevertheless suggestive and stimulated a great deal of additional research. On the basis of this other research, the U.S. surgeon general later ordered that all packages of cigarettes sold in the United States carry the warning that smoking can be hazardous to your health.

7.11 The Effects of Score Transformations

SAT scores and GPAs are not the prettiest numbers. SAT scores are big, up in the hundreds. In fact, if we add the verbal and math SAT scores for each student, as college admissions committees commonly do, the sum is often over 1,000. GPAs, on the other hand, are small numbers with decimal points. Calculated on a four-point scale, where A = 4, B = 3, C = 2, D = 1, and F = 0, GPAs are commonly carried to two decimal places, such as 2.34. However, we can avoid these complications in calculating Pearson's r for SATs and GPAs if we wish.

Suppose the SAT totals are numbers like 1,010, 1,130, and 1,260, all in the 1,000s. We could simplify the figures by subtracting 1,000 from each score. What would change? As we learned in Sections 4.8 and 5.10, the mean would go down by 1,000 points, but because the distances among the scores would remain the same, the standard deviation would not change and neither would the coefficient of correlation between the SAT scores and GPA.

Now let's rid the GPAs of their decimal points by multiplying each by 100; 2.34 becomes 234, for example. The distribution of GPAs thus acquires a mean that is 100 times greater and a standard deviation also 100 times greater. (Again, see Sections 4.8 and 5.10 for the principles.) But r would not change. The general principle is that *r is unaffected by any positive linear transformation of the raw scores*. (Linear transformations, of course, will not change the rankings for Spearman's r_S.) The z scores used in Formula 7.1 are, in fact, linear transformations of the raw scores. A **linear transformation**, you recall, is one in which each raw score changes only by the addition of a constant, the subtraction of a constant, multiplication by a constant, or division by a constant. We can compute positive linear transformations on just the X scores, on just the Y scores, or on both, and r will remain constant.

7.12 Cautions Concerning Correlation Coefficients

You might think that interpreting a correlation coefficient would be straightforward, but that is not always true. We have just seen in Section 7.10, for example, that establishing causation between two correlated variables can be complex. We will now discuss some other difficulties that may arise.

1. *Remember that Pearson's and Spearman's correlation coefficients are appropriate only for linear relationships.* A "hugging" principle is involved: The more closely the scores hug the straight line of best fit, the higher the value of r. What would happen if we calculated r for the curvilinear data graphed in Figure 7.10? It would come out positive, as it ought to, because the data points swarm from lower left to upper right. But r would insist on telling us how closely the points hug a straight line, and because the points do not hug the line very closely, the absolute value of r would be low. The relationship between X and Y here is strong, though, for there is little scatter around the curved line. When data for one or both variables are not linear, other measures of association are better. For more information, see Glass and Hopkins (2008).

range of talent the variability in the distribution of X and Y scores

2. *The correlation coefficient is sensitive to the* ***range of talent*** *(the variability) characterizing the measurements of the two variables.* For example, look at Figure 7.11. If a university had only minimal entrance requirements, the relationship between total SAT scores and freshman GPA might look like that shown in Figure 7.11*a*. Here, SAT scores range from 800 to 1,400 and there is a moderate degree of relationship between the two variables (the cluster of dots slope from the lower left to the upper right). However, suppose that a more selective private university admitted students only with SAT scores of 1,200 or higher. These would be the students represented by the dots to the right of the vertical line in Figure 7.11*a*. The scatter diagram for these individuals would look like that represented by Figure 7.11*b*. As you can see, the relationship between SAT totals and freshman GPAs is much weaker when the range of talent has been restricted. Had we looked only at this restricted group of individuals, we would have underestimated the value of the SAT as a screening device for all applicants who apply to college. It is common to find that the correlation coefficient between scores on an academic aptitude test and academic achievement is highest among grade school children, lower among

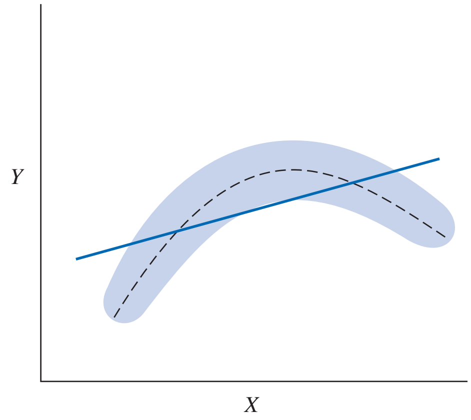


FIGURE 7.10 A curvilinear relationship between X and Y to which a straight line has been fitted. Observations of age (X) and strength of grip (Y) would yield data like those plotted here.

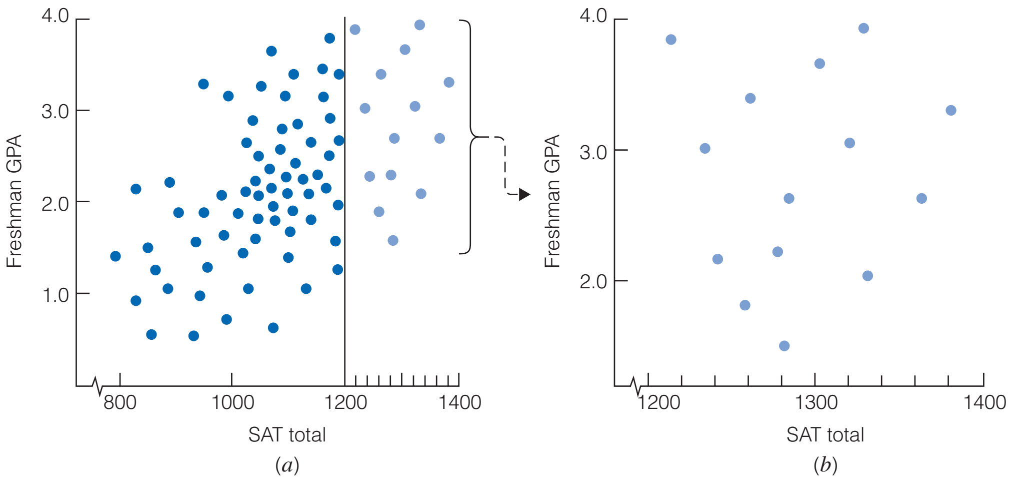


FIGURE 7.11 Relations between SAT scores and freshman GPA when range is unrestricted (*a*) and when it is restricted (*b*).

high school seniors, and still lower among college students. These differences may indicate not that something different is going on among the three groups as much as that the range of ability is greatest among the students in the lower grades and is successively smaller as one progresses into the realms of higher education. Thus, restricting the range, whether in *X*, in *Y*, or both, results in a lower correlation coefficient.

3. *The correlation coefficient tends to be an overestimate in discontinuous distributions*. Suppose someone in the admissions office of a university wants to know the relation between SAT score and freshman GPA for last year's entering class. At lunch with a friend from the student personnel office, he learns that a ready-made sample of personnel records awaits him. His friend is sending notices to those who failed during their first year and is also sending notices to those who earned scholastic honors during their first year. His friend offers to pass on the records of these students, thus saving him the effort of pulling records from the file. If the person from the admissions office accepts the offer, pools the two sets of personnel records, and computes the coefficient of correlation between test score and GPA, the data might look like those pictured in Figure 7.12. Notice that a distribution that usually would be continuous has

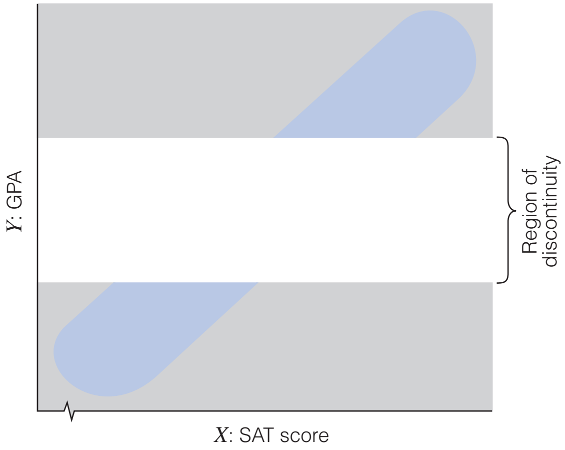


FIGURE 7.12 Scatter diagram for discontinuous data.

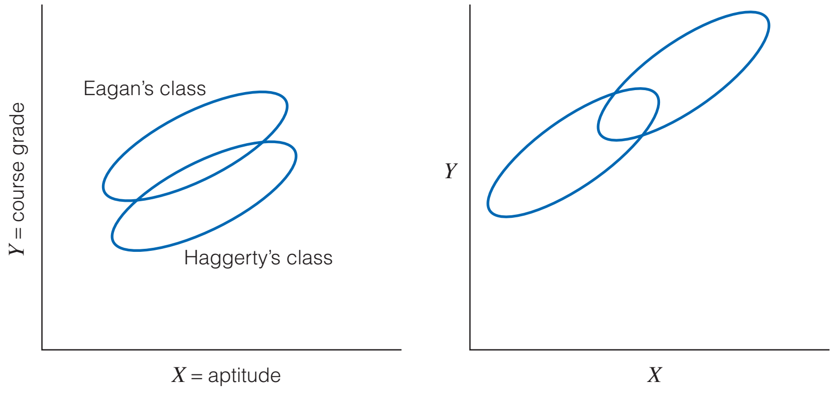
been rendered discontinuous by excluding students with intermediate GPAs. If you take a sample in this manner, it will yield a correlation coefficient different from what you would obtain by sampling *all* elements of the population. *Usually, discontinuity, whether in* X, *in* Y, *or in both, results in a higher correlation coefficient.*

4. *The correlation coefficient, like other statistics, is subject to sampling variation.* Depending on the particular sample, the coefficient may be higher or lower than it would be in a different sample from the same population. It will fluctuate more from sample to sample when the samples are small than when they are large. For very small samples, r is quite unstable from sample to sample, so we cannot trust it to come close to the value for the population.

Once a correlation coefficient is calculated, the next question researchers most frequently ask is whether there is really any correlation at all. That is, can it be that the obtained (sample) coefficient differs from a true (population) coefficient of zero only because of chance (random sampling variation)? We commonly demand that for the sample coefficient to be declared "significantly" different from zero, it must differ enough that such a departure would occur by chance only 5% (or, more rigorously, 1%) of the time if the hypothesis of no correlation in the population were really true. We will treat this question in detail in Chapter 16 after laying the groundwork for statistical inference.

5. *When samples are pooled, the correlation for the combined data depends on where the sample values lie relative to one another in both the* X *and* Y *dimensions.* Suppose Professor Haggerty, a physics instructor, obtains the correlation between academic aptitude test score and grade given in one of her courses; let us say that it is $+.50$. She persuades Professor Eagan, a colleague who teaches the same course, to repeat the study in his class to verify the result. Professor Eagan does so and also finds the correlation to be $+.50$. In a moment of idle curiosity, Haggerty decides to pool the pairs of scores from the two classes and recalculate the correlation coefficient. She does so, and obtains a value of $+.30$! How could this be?

The answer lies in something well known to physics students although perhaps inadequately appreciated by the instructors. The fact of the matter is that the first instructor is known as "Hard-Nosed Haggerty" and the second as "Easy-Aces Eagan." Professor Haggerty is a hard grader. Professor Eagan, on the other hand, thinks all students are wonderful and rarely gives a grade below C. Consequently, the scatter diagrams obtained from the two samples are like those shown in Figure 7.13*a*. Note that the ability of the students (the X distributions) appears to be about the same in the two classes, but the distributions of grades (the Y distributions) are not. The reason for the different correlation coefficient is that when the data are pooled, the points no longer hug the regression line (which now must be a line lying amidst the two distributions) as closely as they did in each separate distribution.



(*a*) (*b*)

FIGURE 7.13 Correlation resulting from the pooling of data from heterogeneous samples.

In the case illustrated, the distributions differed between samples in the mean of Y but not in the mean of X. Other differences are quite possible. Figure 7.13*b* shows a situation in which a second sample differs in that *both* the mean of X and the mean of Y are higher than in the first sample. In this case, the correlation will therefore be greater among the pooled data than among the separate samples.

6. *There is no such thing as "the" correlation coefficient for two variables.* The exact value of r or r_S depends on several things. One is the particular sample of subjects we happen to get, as just noted. Another is exactly how we measure the variables; swap the SAT for a different aptitude test, and the correlation with GPA will change. A third factor is the specific conditions under which the variables are operating. If your college changes its grading policies—allowing students to take more courses on a pass-fail basis, for example—the correlation between SAT and GPA may also change.

Any report of r or r_S values, therefore, should include a careful description of the measures used and the circumstances under which the correlation coefficients were obtained. Similarly, values reported by others may or may not hold up under the conditions with which you are concerned. Take the others' results only as working hypotheses, subject to confirmation under your conditions.

7.13 Summary

In behavioral science, it is often important to determine the degree to which two variables vary together. The two basic tasks are (1) determining the degree of association between two variables and (2) predicting a person's standing in one variable when we know the standing in an associated variable. In this chapter, we discussed the problem of association; we consider prediction in the next chapter.

The Pearson product-moment correlation coefficient, r, is the most commonly used measure of association when two quantitative variables are characterized by a linear (straight-line) relationship. It reflects agreement between relative standing in one variable and relative standing in the other and so is an index of how well bivariate data points hug the straight line of best fit. The sign of the coefficient specifies whether the two variables are positively or negatively (inversely) related. The magnitude of the coefficient varies from zero when no association is present to ± 1.00 when the correlation is perfect.

Making a scatter diagram of the data points of a bivariate distribution is an excellent preliminary step to see whether there is any association between two variables. It also provides a rough check on the accuracy of the calculated value of r, and it lets us examine several conditions (such as the linearity of the relationship) that may influence the correlation coefficient and its interpretation.

We can use Spearman's correlation coefficient, r_S, in place of Pearson's r if the paired observations are converted to ranks. In fact, Spearman's correlation is found merely by applying Pearson's procedure to the ranks. However, there should be no ties in ranks, so we typically use r_S only in circumstances in which the sample size is rather small.

A degree of correlation between two variables means that they vary together to some extent, but it does not necessarily mean that a causal relationship exists between them. We must also keep in mind that if Pearson's r is applied to data that show a curvilinear relationship, the index obtained will underestimate the degree of relationship. Other techniques exist for estimating relationships when two variables are nonlinearly related.

There is no such thing as *the* correlation between two variables. Many factors may influence the correlation coefficient's value. It is sensitive to the range of talent (variability) characterizing the measurements of the two variables and also the continuity of the distributions. Like other statistics, it is also subject to sampling variation. Variation from sample to sample will, as usual, be greater when samples are small than when they are large. The value of r also depends on how the variables were measured and on the conditions under which the measurements were obtained. We shall therefore need careful description of these circumstances to interpret *r* properly.

Mathematical Notes

Note 7.1 Derivation of Deviation-Score Formula for *r* from *z*-Score Formula (*Ref.*: Section 7.6)

$$
\begin{aligned}
r &= \frac{\sum (z_X z_Y)}{n} \\
&= \frac{\sum \left(\dfrac{X-\overline{X}}{S_X}\right)\left(\dfrac{Y-\overline{Y}}{S_Y}\right)}{n} \\
&= \left[\sum \left(\dfrac{X-\overline{X}}{S_X}\right)\left(\dfrac{Y-\overline{Y}}{S_Y}\right)\right]\frac{1}{n} \\
&= \frac{\sum (X-\overline{X})(Y-\overline{Y})}{nS_X S_Y}
\end{aligned}
$$

Note 7.2 The Raw-Score Equivalent of $\Sigma(X-\overline{X})(Y-\overline{Y})$ (*Ref.*: Section 7.7)

$$
\begin{aligned}
\sum (X-\overline{X})(Y-\overline{Y}) &= \sum (XY - \overline{X}Y - \overline{Y}X + \overline{X}\,\overline{Y}) \\
&= \sum XY - \overline{X}\sum Y - \overline{Y}\sum X + n\overline{X}\,\overline{Y} \\
&= \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} - \frac{\left(\sum Y\right)\left(\sum X\right)}{n} + \frac{\cancel{n}\left(\sum X\right)\left(\sum Y\right)}{(\cancel{n})(n)} \\
&= \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}
\end{aligned}
$$

Key Terms, Concepts, and Symbols

correlation (p. 98)
prediction (p. 98)
Sir Francis Galton (p. 98)
bivariate distribution (p. 99)
Karl Pearson (p. 99)
scatter diagram (p. 99)
linear relationship (p. 101)
curvilinear relationship (p. 102)
positive correlation (p. 102)
negative correlation (p. 103)
no correlation (p. 103)
coefficient of correlation (p. 103)
r (p. 103)
product-moment correlation coefficient (p. 106)
cross-products (p. 106)
Spearman's rank-order correlation coefficient (p. 112)
r_S (p. 112)
linear transformation (p. 115)
range of talent (p. 116)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Seven students made the following scores on two quizzes, X and Y:

Student:	A	B	C	D	E	F	G
Score on X:	3	9	7	8	4	6	7
Score on Y:	2	7	8	6	6	5	7

(a) Construct a scatter diagram of the relationship between these two variables. What is the direction of the relationship? (b) Compute r to two decimals using the raw-score method. (c) If the two quizzes covered the same subject matter, what is it about the students' performance that r measures?

2. Same as Problem 1a–1b, except substitute the following data:

Student:	H	I	J	K	L	M
Score on X:	7	2	1	6	5	3
Score on Y:	7	6	12	4	3	10

3. Using a 10-point scale, two managers independently rate the performance of the same five salespeople.

Salesperson:	H	I	J	K	L
Rating by first manager (X):	9	4	5	3	5
Rating by second manager (Y):	4	8	4	8	7

(a) Construct a scatter diagram and determine the direction of the relationship. (b) Compute r, using the raw-score method. (c) What aspect of the managers' judgments does r measure?

4. In Figure 7.3, one of the scatter diagrams is described by $r = +.68$. The 10 pairs of scores are

X:	30	34	35	36	39	39	40	40	42	46
Y:	72	70	76	80	73	79	76	83	85	81

(a) Verify, by the raw-score method, that $r = +.68$. (b) Suppose that X is a person's score on a personality test designed to measure sociability and Y is the rating of that person's sociability made by a close friend. What do the results say about the *validity* of the personality test (the degree to which it measures what it's supposed to)?

5. Compare Formula 7.4 for calculating $\Sigma(X - \overline{X})\,(Y - \overline{Y})$ from the raw scores X and Y to the formula for calculating $\Sigma(X - \overline{X})^2$ or $\Sigma(Y - \overline{Y})^2$ from the corresponding raw scores X or Y. (The latter equation appears just above Formula 7.4 in Section 7.7.) Formula 7.4 parallels the other formula. How?

6. In most of the examples in this chapter, the variables (e.g., SAT score, GPA) characterize an individual person. However, when computing correlations, the "subject" need not be individuals. Data 7A below gives some of the evidence on smoking and lung cancer (U.S. Department of Health and Human Services, 1982):

DATA 7A

NATION	PER CAPITA CONSUMPTION OF CIGARETTES IN 1930	MALE DEATH RATE (PER MILLION) FROM LUNG CANCER IN 1950
Iceland	240	60
Norway	250	90
Sweden	310	120
Denmark	370	160
Australia	450	160
Holland	450	240
Canada	500	150
Switzerland	530	250
Finland	1,110	350
Great Britain	1,130	460
United States	1,280	190

(a) In Data 7A, the "subjects" of the research are not individuals, but rather what? (b) Make a scatter diagram of these data. (c) In words, describe the relationship between the two variables. (d) Estimate the correlation coefficient. (e) Calculate *r*.

7. Rank each of the following sets of scores from lowest to highest and perform the check suggested in Section 7.9: (a) 11, 13, 15, 19, 19; (b) 11, 13, 13, 13, 15, 19, 19, 21.

8. Two judges at an art show are asked to place the six pictures that reached the "finals" in order of merit. Their rankings are

Picture:	A	B	C	D	E	F
Judge A:	3	2	6	4	1	5
Judge B:	2	1	5	3	4	6

Calculate r_S, the Spearman rank-order coefficient of correlation, as an index of agreement between the two judges.

9. For Problem 8, if you calculated Pearson's *r* as the index of agreement between the two judges, would you expect *r* to be identical with r_S? Explain.

10. (a) Consider the data for Problem 6 about the relationship between smoking and lung cancer. Translate these scores to ranks and compute r_S, the Spearman coefficient. (b) Pearson's *r* for these data is +.74. Does this agree with the r_S you calculated? Explain.

11. In what sense do linear transformations involve a straight line?

12. An *r* of +.60 was obtained between points earned on a spelling test and IQ for all current members of the sixth grade in a particular school. For each of the following, state whether the correlation would be affected, and if so, how. If it would be affected in an unpredictable manner, say so. Treat each question as independent of the others. (a) Score in spelling is changed to number of answers not correct. (b) Each IQ is divided by 10. (c) Ten points are added to each spelling score. (d) Ten points are added to each spelling score *and* each IQ is divided by 10. (e) Spelling scores are converted to *z* scores. (f) Spelling scores are converted to *z* scores and IQ to standard scores with a mean of 50 and a standard deviation of 10.

13. The more cigarettes a person has smoked over his or her lifetime, the greater the probability that the person has developed lung cancer. (a) What is the direction of the correlation between these two variables? (b) This finding does not, by itself, prove that smoking causes lung cancer. Why not? (c) If smoking does not cause lung cancer, what might explain the correlation between the two variables? (d) Medical authorities believe that a cause-and-effect relationship between smoking and lung cancer has been proven well beyond reasonable doubt. What kind of evidence do you think has persuaded them?

14. Across the cities of Europe, the greater the number of storks nesting within the city limits, the greater the number of babies born in a given year. (a) What is the direction of this correlation? (b) Does this correlation prove that storks bring babies? Why or why not? (c) Explain the correlation without assuming that birds affect birthrates. (d) Which part of Figure 7.9 does your explanation illustrate?

15. Over the 365 days of any year in any major city, the greater the quantity of ice cream sold, the greater the number of murders. (a) What is the direction of this correlation? (b) Does this correlation prove that eating ice cream drives people to violence? Why or why not? (c) Explain the correlation without assuming a causal connection between the two variables. (d) Which part of Figure 7.9 does your explanation illustrate?

16. Does a low value of *r* between two variables necessarily indicate that there is very little association between them? Explain.

17. The personnel director for the Miniscule Corporation often lunches with her counterpart for the Megasize Corporation. This noon she mentions her difficulty in selecting workers for an assembly line. Her friend says that she's had good results with the Acme Aptitude Test: Among 400 workers, score on the test correlated .77 with performance on the job. What caution might you offer the Minuscule director before she rushes out to buy the test?

18. Among a group of mentally handicapped 10-year-old children, it is found that the correlation between IQ and reading achievement is +.25. However, on a school-wide basis the correlation is +.50. Assuming these results to be typical, what explanation do you suggest?

19. To study creativity, a number of outstandingly creative persons are assembled and tested. It is found that ratings on creativity and IQ correlate +.20. It is concluded that creativity is really quite different from intelligence. Describe any objections you have to this conclusion.

20. One way to learn whether an aptitude test works is to test applicants for the particular job, hire them all, and subsequently correlate test score with a measure of job proficiency. Another way, usually more practical, is to test all currently employed workers of the particular type and obtain the same correlation. Is there any reason to think that the coefficient might be different when obtained in these two ways? Explain.

21. For which group would you expect a higher correlation between height and weight: all students at a 4-year high school, or a sample composed (only) of the ninth graders and the seniors at the same school? Explain.

22. **Calculating *r* with IBM SPSS.** To learn how to use IBM SPSS, you do not need large data sets. You have already calculated *r* by the raw-score method in Problems 1, 2, and 3. For practice, calculate *r* with IBM SPSS. Once you feel comfortable with the software, calculate *r* for the following problem.

In numerous studies, psychologists and sociologists have found a positive correlation between exposure to violence in the media (e.g., television, movies) and acts of aggression toward others. (See Anderson et al., 2003; Christenson & Wood, 2006; Johnson et al., 2006; for a review). Suppose that you replicate the study with fifty 16-years-old boys and measure the mean number of media exposures per week (X) and acts of aggression (mild to severe) during the past year (Y). Your results are as follows. What is *r*?

Subject	X	Y
1	3	9
2	15	8
3	35	13
4	25	11
5	20	18
6	37	23
7	9	14
8	5	8
9	27	11
10	17	10

Subject	X	Y
11	7	13
12	26	21
13	40	7
14	30	20
15	21	9
16	2	2
17	29	5
18	32	15
19	12	9
20	18	12

Subject	X	Y
21	30	11
22	26	13
23	20	5
24	35	7
25	4	4
26	13	10
27	15	5
28	28	14
29	36	26
30	22	11

Subject	X	Y
31	7	8
32	10	10
33	34	11
34	32	11
35	20	20
36	10	2
37	23	14
38	40	18
39	35	19
40	30	16

Subject	X	Y
41	5	12
42	10	8
43	38	14
44	27	9
45	15	12
46	15	17
47	25	7
48	25	16
49	30	15
50	20	7

CHAPTER 8

Prediction

When you have finished studying this chapter, you should be able to:

- Understand that the greater the linear correlation between two variables, the better we are able to predict the standing in one variable from knowledge of the standing in the other;
- Understand that to predict Y from X, we first determine a line of best fit—the regression line—that estimates the mean of Y for each value of X;
- Plot the regression line according to the least-squares principle;
- Predict Y from X by the raw-score form of the regression equation;
- Understand that except in the case of perfect correlations (±1.00), the predicted value of Y from X will be in error;
- Calculate the standard error of estimate, S_{YX};
- Use IBM SPSS to calculate a regression equation and S_{YX};
- Explain the underlying assumptions (linear relationship, homoscedasticity, normal distribution of scores) in predicting Y from X; and
- Understand that predicting Y from X or X from Y does not prove causation.

In Chapter 7, you learned that SAT scores are positively correlated with freshman-year grade point averages (Slack & Porter, 1980). If someone scores well above average on the SATs, there is a good chance that his or her GPA is also above average. In general, if two variables are correlated, it is possible to make a **prediction**, with better than chance accuracy, of the standing in one of the variables from knowledge of the standing in the other. *The closer the relationship between the two variables, the higher the correlation coefficient and the better the prediction.* Still, the value of the coefficient by itself does not tell us how to proceed. How, then, do we make predictions from correlated data?

prediction when two variables are correlated, estimating the value of one variable from the known value of the other; the better the correlation between the variables, the better the prediction

In this chapter we consider predictions only for cases in which the relationship between bivariate data can best be described by a straight line. The Pearson correlation coefficient was developed for linear relationships. Here we consider the equation for the straight line of best fit and learn how to make predictions. If the correlation between two variables is less than perfect, as it always is in real-world cases, then our prediction will also be less than perfect. This means that we will make mistakes when making predictions. Thus, we must also consider the error of prediction and how to measure it.

8.1 The Problem of Prediction

Given the correlation between the SAT and freshman-year GPA, college admissions committees looking at applicants' SAT scores can predict with some success how well those students would do if admitted. Two possible ways to make that prediction are demonstrated with the scatter diagram in Figure 8.1. In the scatter diagram, we can see the positive association between

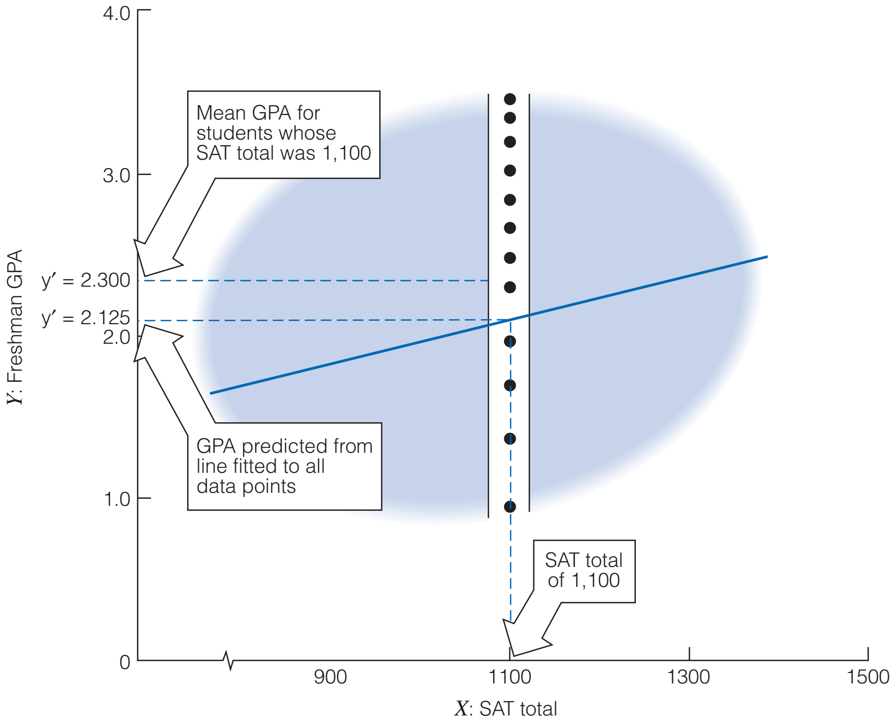


FIGURE 8.1 Prediction of Y for persons with SAT total scores of 1,100 from column mean and from line of best fit to Y values.

SAT score and freshman-year GPA for a sample of the students admitted to a hypothetical college in the past. Consider a new applicant with an SAT total of 1,100. To take the simplest possible approach to predict his or her GPA, we could look only at the students who had that particular score. Their data points appear in the column erected over the value 1,100 on the abscissa. Suppose the mean GPA for those students was 2.3. Then, 2.3 will be our prediction for this applicant.

This method of prediction has a major shortcoming; it ignores cases that have SAT scores other than 1,100. The prediction is based on a relatively small sample and is therefore somewhat unstable; the prediction from another sample of students with scores of 1,100 may differ markedly from this one. However, in Chapter 11 you will learn that larger samples from a given population vary less among themselves than do smaller samples. Thus, if we can find a way to use the full sample of observations, we can generate predictions that are more resistant to sampling variation.

regression line in prediction of *Y* from *X*, the straight line of best fit to the *Y* values

regression equation the equation that locates the regression line

If it is reasonable to assume that *X* and *Y* are linearly related (our scatter diagram should tell us), we can improve our prediction of *Y* from *X* by finding the straight line of best fit to the *Y* values. This will be a line determined by all the scores in the sample on hand. Statisticians call the line of best fit a **regression line**; the equation that locates the line is the **regression equation**. The regression line is shown in Figure 8.1, and as shown, we can use it to make predictions for new cases. Just start with the SAT score on the abscissa, go up to the regression line, and then go over to the ordinate. By this method, the best prediction for a student with an SAT total of 1,100 is a GPA of about 2.12.

Predictions made by this technique are better in their resistance to sampling variation, but two limitations remain. First, a regression line that has been fitted to a sample of points is probably not the same as the line that would best fit the entire population. (Other things

"What's even more astonishing is it coincides exactly with the World Series."

© The New Yorker Collection 1984 James Stevenson from cartoonbank.com. All Rights Reserved.

being equal, however, the larger the sample, the closer the approximation.) Second, the technique depends on the assumption that a straight line is a reasonable description of the interrelationship between X and Y. Fortunately, the **assumption of linearity** is often satisfactory. Let us now consider how to define "best fit."

assumption of linearity the assumption that a straight line is the best description of the relationship between X and Y

8.2 The Criterion of Best Fit

How do we find the straight line of best fit? Best fit could be defined in more than one way. Karl Pearson's solution to this problem was to apply the *least-squares criterion*. Figure 8.2 illustrates his thinking for predicting Y from X.

Y' symbol for the predicted value of Y

The figure shows a bivariate distribution and a possible regression line. How good are the predictions from this line? For the seven cases shown in the scatter diagram, the errors of prediction appear as vertical lines, each connecting the actual value of Y to the predicted value, which we call **Y'**, given by the regression line. The longer the vertical line, the greater the error in prediction.

d_Y symbol for the discrepancy of $Y - Y'$

criterion for the regression line the **line of best fit** is the one that minimizes the sum of the squares of the discrepancies between the actual value of Y and the predicted value (called the **least-squares criterion**)

Let **d_Y** stand for the discrepancy between the actual value of Y and the predicted value: $d_Y = Y - Y'$. Pearson's **least-squares criterion** for the regression line is this: the **line of best fit** is the one that minimizes the sum of the squares of the discrepancies. Thus, Σd_Y^2 must be as small as possible.

Why not just minimize the sum of the absolute magnitudes of the discrepancies rather than the sum of the squares? The answer has two parts: (1) It is difficult to deal mathematically with the absolute discrepancies, whereas the squared discrepancies permit mathematical developments of practical value. (2) The location of the regression line and the value of the correlation coefficient will fluctuate less under the influence of random sampling than would happen if another criterion were used.

This is not the first time we have encountered a sum of squared discrepancies. Recall that the sum of the squared deviations from the mean, $\Sigma(X - \overline{X})^2$, is minimal (see Section 5.8 and Note 5.2 at the end of Chapter 5). Is there some connection between Pearson's regression line and the mean? Yes.

First, just as the mean is a least-squares solution to the problem of finding a measure of central tendency, so the regression line is a least-squares solution to the problem of finding the best-fitting straight line. Both minimize the sum of squared discrepancies; thus, they have analogous properties, including resistance to sampling variation.

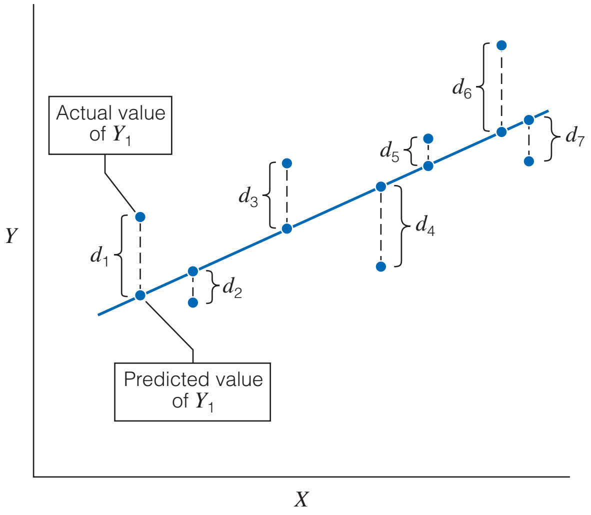


FIGURE 8.2 Discrepancies between seven Y values and the line of regression of Y on X.

Point of Controversy

Least-Squares Regression versus the Resistant Line

In Chapter 4 we discussed why, in most instances, the mean is the best indicator of central tendency. One of the main reasons, you recall, is that the mean is affected by all the scores in the distribution. However, for the same reason, the mean is *not* the best indicator of central tendency when the distribution is severely skewed or there are a few very aberrant scores ("outliers"). In the latter case, the median is often preferred because its value is not affected by outliers.

The regression line is a running mean, a series of means (Section 8.2). It is located such that it minimizes the sum of the squares of the discrepancies between the actual value of Y and the predicted value, Y'. As a score becomes more deviant from the other scores, it increases the sum of the squared deviations at an increasing rate. Thus, an outlier has considerable influence on the location of the line. In other words, the regression line tends to reflect the influence of outliers. (It lacks resistance to the substantial influence of atypical scores.)

resistant line uses the median, rather than the mean, to fit lines to points on a scatter diagram

Because the median is resistant to outliers, John Tukey (1977) and other advocates of exploratory data analysis prefer it for fitting lines to points on a scatter plot (e.g., see Hartwig & Dearing, 1979, and Hoaglin, Mosteller, & Tukey, 2000). The result is called the **resistant line**. The technique for fitting such a line can be found elsewhere (see the cited references).

Which is better, the least-squares regression line or the resistant line? When a distribution contains clear outliers, the resistant line may well be superior. On the other hand, the median, unlike the mean, does not lend itself to further mathematical manipulations. If we use the standard regression equation, further analysis is possible. For example, the standard error of estimate and its applications depend on the fact that we have the leastsquares regression line (see Section 8.5). Tests are available for confirming the presence of outliers in a distribution, and some investigators prefer to make the regression line more resistant by trimming them rather than using a technique (the resistant line) that does not allow for further analysis.

Whichever technique you choose to analyze your own data, remember the following: Never compute and interpret correlation coefficients or slopes and intercepts without *first* plotting the X and Y variables, checking for nonlinearity and outliers, and generally inspecting your data.

Second, *the regression line is actually a kind of mean. It is a running mean, a series of means*. For each value of X, the regression line tells us the mean, or expected, value of Y. In other words, whereas $\overline{Y}$ is the mean of all Y values in a set of scores, Y' is an estimate of the mean of Y given the condition that X has a particular value.

8.3 The Regression Equation: Standard-Score Form

$Y = bX$, or
$Y = bX + a$
general formulas for a straight line (compare to Formulas 8.1 and 8.2)

In practice, we do not need to experiment with various lines on a scatter diagram to find the one that fits best by Pearson's criterion. Recall from your college algebra course that the formula for a straight line is **$Y = bX$**. Here, b is a constant that multiplies each value of X. A straight line may also be expressed as **$Y = bX + a$**, where both a and b are constants. In this section and the next, it will help to notice the similarity between these two formulas for a straight line and Formulas 8.1 and 8.2. We will return to a comparison of the regression equation and that for a straight line in Section 9.2.

The straight line established according to the least-squares principle is the one with this elegantly simple **regression equation**:

$z'_Y = rz_X$
regression equation of Y on X, standard-score formula

REGRESSION OF Y ON X: STANDARD-SCORE FORMULA

$$z'_Y = rz_X \qquad (8.1)$$

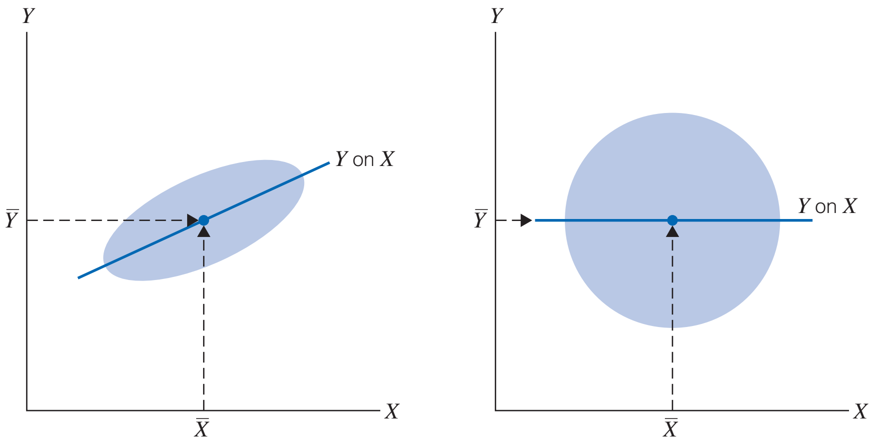
where: z'_Y = the predicted standard-score value of Y
r = the coefficient of correlation between X and Y
z_X = the standard-score value of X from which z'_Y is predicted

Notice the title of this equation; it is customary to speak of the **regression of Y on X** when we predict Y from X.

z'_Y
symbol for the predicted standard-score value of Y

The standard-score regression equation is seldom used for actual predictions, because the scores from which a prediction is made are usually in the form of raw scores. However, it reveals some important things about the workings of prediction. First, suppose we wish to predict Y from a case at the mean of X. Because the mean of a set of scores expressed in standard-score form is always zero (see Section 5.11), the predicted standard score on Y is also zero: $z'_Y = (r)(0) = 0$, the mean of Y, no matter what the value of r. Thus, *for all values of* r, *the regression equation predicts that a case at the mean of* X *will score at the mean of* Y. On a graph, then, the regression line will always pass through the point defined by the mean of X and the mean of Y, as illustrated in Figure 8.3.

Second, if $r = 0$, then the predicted standard-score value of Y will always be zero: $z'_Y = (0)(z_X) = 0$. In raw-score terms, *if the correlation is zero, then the predicted value of* Y *is the mean of* Y *no matter what value of* X *is used to predict* Y. The implications are both interesting and logical. If knowing the value of X affords no advantage in predicting Y, what value of Y shall we predict? Not only is the mean of Y an intuitively reasonable prediction, it satisfies the least-squares criterion as well; the sum of the squares of errors of prediction is minimized. Figure 8.3*b* shows the line of regression of Y on X when $r = 0$. We will discuss these topics in greater detail in the next chapter (Sections 9.2–9.3).



(*a*) (*b*)

FIGURE 8.3 Two scatter diagrams showing the lines of regression of Y on X when (*a*) $r > 0$ and (*b*) $r = 0$. Both regression lines pass through the point $X = \overline{X}$, $Y = \overline{Y}$.

8.4 The Regression Equation: Raw-Score Form

Although the standard-score formula helps us understand the straight line of best fit, most practical problems in prediction are expressed in terms of raw scores. A little algebra transforms the standard-score form of the regression equation into a raw-score form that is more useful for practical work. First, replace the *z* scores with their raw-score equivalents:

$$\frac{Y' - \overline{Y}}{S_Y} = r\,\frac{X - \overline{X}}{S_X}$$

Then, solving the equation for Y', we obtain the raw-score form of the regression equation:

$$\mathbf{Y' = \left(r\,\frac{S_Y}{S_X}\right)X - \left(r\,\frac{S_Y}{S_X}\right)\overline{X} + \overline{Y}}$$

regression equation of *Y* on *X*, raw-score formula

REGRESSION OF *Y* ON *X*:
RAW-SCORE FORMULA

$$Y' = \left(r\,\frac{S_Y}{S_X}\right)X - \left(r\,\frac{S_Y}{S_X}\right)\overline{X} + \overline{Y} \tag{8.2}$$

where: Y' = the predicted raw score in Y
S_X and S_Y = the two standard deviations
$\overline{X}$ and $\overline{Y}$ = the two means
r = the correlation coefficient between X and Y

The formula looks more complicated than it really is. To predict Y', we need to know the value of X from which the particular prediction is to be made, the two means, the two standard deviations, and the value of the correlation coefficient. For example, let us predict the freshman-year GPA for an applicant to a college where past experience has yielded the following figures:

Given

X = SAT total
Y = GPA
$\overline{X} = 1{,}000$
$\overline{Y} = 2.00$
$S_X = 140.0$
$S_Y = .50$
$r = +.35$

Problem An applicant's SAT scores total 1,100. What is the best prediction of the person's freshman-year GPA if he or she is admitted to the college?

Solution **Step 1:** Insert the means, standard deviations, and the correlation coefficient in the raw-score formula and simplify it:

$$\begin{aligned} Y' &= \left(r\,\frac{S_Y}{S_X}\right)X - \left(r\,\frac{S_Y}{S_X}\right)\overline{X} + \overline{Y} \\ &= (+.35)\left(\frac{.50}{140.0}\right)X - (.35)\left(\frac{.50}{140.0}\right)1{,}000 + 2.00 \\ &= .00125X - 1.250 + 2.00 \\ &= .00125X + .75 \end{aligned}$$

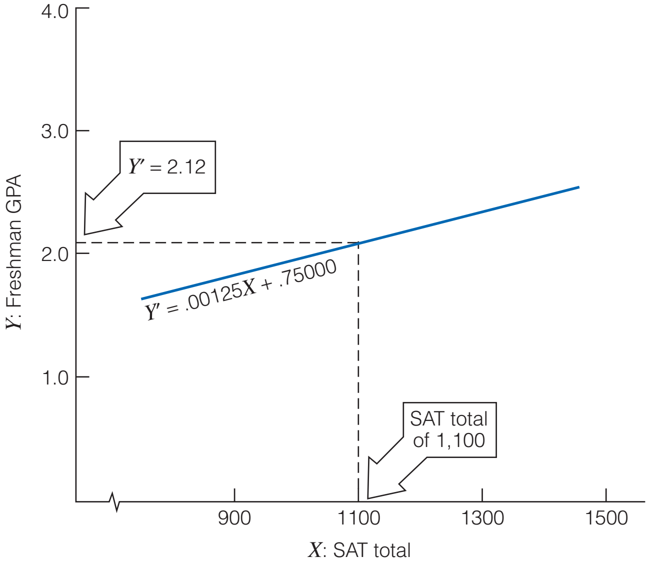


FIGURE 8.4 Graph of regression equation for predicting freshman GPA from SAT total. Data from Section 8.4.

Step 2: Insert the value of X from which the prediction is to be made, and find the predicted value Y':

$$\begin{aligned} Y' &= .00125X + .75 \\ &= (.00125)(1{,}100) + .75 \\ &= 1.375 + .75 \\ &= 2.12 \end{aligned}$$

Notice that we did not insert the particular value of X from which Y' is to be predicted until after we simplified the equation. This is particularly desirable if we wish to make more than one prediction, as is usually the case. For example, if we now wish to predict the freshman-year GPA for a person whose SAT total is 1,200, just insert that value in the formula $Y' = .00125X + .75$ to arrive at a prediction of 2.25 for the GPA.

Because the regression equation specifies a straight line, it is easy to graph the equation and then use the line to make predictions from various values of X. Figure 8.4 shows the regression line for the preceding example.

8.5 Error of Prediction: The Standard Error of Estimate

The regression equation states what value of Y is expected (Y') when X takes a particular value. Of course, Y' is not likely to be the actual value of Y that corresponds to the particular value of X. If a man is 6 feet tall, the appropriate regression equation may predict his weight to be 175 lb., but we do not expect a given 6-footer to have exactly that weight. Remember, *the predicted value is only an estimate of the mean value of* Y *for cases with the given score on* X. Thus, 175 lb. is only a "best estimate" of our 6-footer's weight. If the correlation is low, expect considerable variation of actual values around the predicted value. If the correlation is high, the actual values will cluster more closely about the predicted value. *Only when the correlation is unity* (±1.00) *will the actual values regularly and precisely equal the predicted values.*

What we need is a way to measure the predictive error, the variability of the actual Y values around the predicted value (Y'). Such a measure will have desirable properties if it is cast in

standard error of estimate a measure of variability of the actual Y values around the predicted value, Y'

$\mathbf{S_{YX}}$ symbol for the standard error of estimate

$$S_{YX} = \sqrt{\frac{\Sigma(Y - Y')^2}{n}}$$

formula for the standard error of estimate

the form of a standard deviation (see Chapter 5). The **standard error of estimate**, S_{YX}, is exactly that kind of measure. Compare its formula with that of the standard deviation:

STANDARD DEVIATION OF *Y*

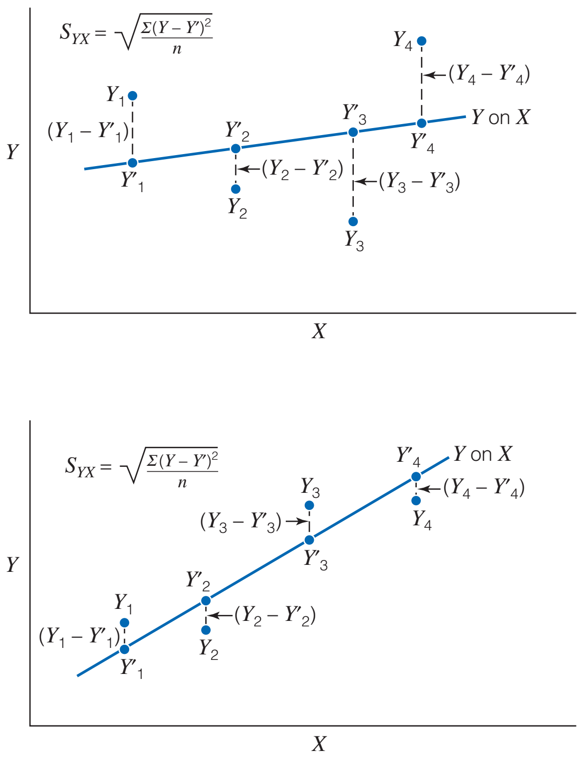
$$S_Y = \sqrt{\frac{\sum(Y - \overline{Y})^2}{n}}$$

STANDARD ERROR OF ESTIMATE OF *Y* ON *X*

$$S_{YX} = \sqrt{\frac{\sum(Y - Y')^2}{n}} \qquad (8.3)$$

Figure 8.5 illustrates the discrepancies, $(Y - Y')$, on which Formula 8.3 is based. These are the same values we called d_Y in Section 8.2 and illustrated in Figure 8.2.

When the correlation is perfect, every value of $(Y - Y')$ is zero, and therefore S_{YX} is zero. In short, there is no error of prediction. When the correlation is zero, $Y' = \overline{Y}$ for all values


$S_{YX} = \sqrt{\frac{\Sigma(Y - Y')^2}{n}}$
Y_1
$(Y_1 - Y'_1)$
Y'_1
Y'_2
$(Y_2 - Y'_2)$
Y_2
Y'_3
$(Y_3 - Y'_3)$
Y_3
Y_4
$(Y_4 - Y'_4)$
Y'_4
Y on X
Y
X
$S_{YX} = \sqrt{\frac{\Sigma(Y - Y')^2}{n}}$
Y'_4
Y on X
$(Y_4 - Y'_4)$
Y_4
Y_3
$(Y_3 - Y'_3)$
Y'_3
Y'_2
$(Y_2 - Y'_2)$
Y_2
Y_1
$(Y_1 - Y'_1)$
Y'_1
Y
X

These large values of $(Y - Y')$ are characteristic of a low correlation; they will lead to a large value of S_{YX}.

These small values of $(Y - Y')$ are characteristic of a high correlation; they will lead to a small value of S_{YX}.

FIGURE 8.5 The standard error of estimate (S_{YX}) as a function of the magnitude of the $(Y - Y')$ discrepancies.

TABLE 8.1 *Calculation of* S_Y *and* S_{YX} *(Data from Table 7.2)*

SUBJECT	① X	② Y	③ $(Y-\overline{Y})$	④ $(Y-\overline{Y})^2$	⑤ Y'	⑥ $(Y-Y')$	⑦ $(Y-Y')^2$
A	17	9	−4.4	19.36	13.71	−4.71	22.1841
B	8	13	−0.4	0.16	6.67	6.33	40.0689
C	8	7	−6.4	40.96	6.67	0.33	0.1089
D	20	18	+4.6	21.16	16.06	1.94	3.7636
E	14	11	−2.4	5.76	11.36	−0.36	0.1296
F	7	2	−11.4	129.96	5.88	−3.88	15.0544
G	21	5	−8.4	70.56	16.85	−11.85	140.4225
H	22	15	+1.6	2.56	17.63	−2.63	6.9169
I	19	26	+12.6	158.76	15.28	10.72	114.9184
J	30	28	+14.6	213.16	23.89	4.11	16.8921
$\Sigma =$	166	134	0.0	662.40	134.00	0.00	360.4594
$\Sigma/n =$	16.6	13.4		66.24	13.40		36.0459

Calculation of S_Y:

$$S_Y = \sqrt{\Sigma(Y-\overline{Y})^2/n}$$

$$= \sqrt{66.24}$$

$$= 8.139$$

Calculation of S_{YX}:

$$S_{YX} = \sqrt{(Y-Y')^2/n}$$

$$= \sqrt{36.0459}$$

$$= 6.00$$

of X (as noted in Section 8.3). We may therefore substitute $\overline{Y}$ for Y' in Formula 8.3 to obtain $\sqrt{\Sigma(Y-\overline{Y})^2/n}$, which is S_Y. So, we can see that *the value of* S_{YX} *ranges from zero when the correlation is perfect to* S_Y *when there is no correlation at all.*

We will illustrate some of these characteristics using the 10 pairs of scores from Table 7.2. In practice, 10 cases would be too small a sample to yield reliable results in the application of the regression equation and the standard error of estimate (see Section 8.8). For present purposes, however, it will help keep things simple.

Table 8.1 shows the calculation of S_Y and S_{YX} for this small sample. The first two columns list the paired raw scores. Columns 3 and 4 contain the values needed to calculate S_Y. Column 5 presents the value of Y predicted by the regression equation for each value of X. Columns 6 and 7 contain the values needed for the calculation of S_{YX}. Because the correlation coefficient here (+.675) lies between zero and one, we expect S_{YX} to be less than S_Y (8.139) but greater than zero. In fact, it is 6.00.

Is S_{YX} really a standard deviation? Formula 8.3 looks very much like the formula for the standard deviation, except that Y' rather than $\overline{Y}$ is subtracted from each actual value of Y. But as we noted earlier (Section 8.2), Y' is a kind of mean, namely, the estimated mean of Y when X has a particular value. Therefore, S_{YX} is a kind of standard deviation; S_{YX} *is the standard deviation of obtained* Y *scores about the predicted* Y *scores.*

Because S_{YX} really is a standard deviation, it should have the properties of one. Recall that one property is that the sum of the squares of the deviations of each score from the mean, $\Sigma(Y-\overline{Y})^2$, is a minimum (Section 5.8). This is also true for the standard error of estimate: The regression line is laid down so as to minimize $\Sigma(Y-Y')^2$. A second property of the standard deviation that S_{YX} shares is that the sum of the deviations of the scores (unsquared) from the mean of the scores must be zero (see Section 4.5). Look at the bottoms of Columns 3 and 6 in Table 8.1. Note 8.1 at the end of the chapter shows this property to be a general one, not just a coincidence.

8.6 An Alternative (and Preferred) Formula for S_{YX}

The formula we gave in the previous section for the standard error of estimate, $S_{YX} = \sqrt{\Sigma(Y - Y')^2/n}$, is good for understanding the statistic. However, for actual prediction, it is better to use Formula 8.4:

$S_{YX} = S_Y\sqrt{1 - r^2}$ preferred formula for the standard error of estimate

STANDARD ERROR OF
ESTIMATE OF *Y* ON *X*

$$S_{YX} = S_Y\sqrt{1 - r^2} \qquad (8.4)$$

Formula 8.4 is the exact algebraic equivalent of Formula 8.3. Notice, for example, that if $r = 0$, then $S_{YX} = S_Y$, and that if $r = +1.00$, then $S_{YX} = 0$. We can use Formula 8.4 to calculate S_{YX} for the data of Tables 7.2 and 8.1,

$$\begin{aligned} S_{YX} &= 8.139\ \sqrt{1 - (.675)^2} \\ &= 8.139\ \sqrt{1 - .456} \\ &= 8.139\ \sqrt{.544} \\ &= (8.139)(.738) \\ &= 6.01 \end{aligned}$$

which differs from the answer obtained by Formula 8.3 (see Table 8.1) only because of rounding error.

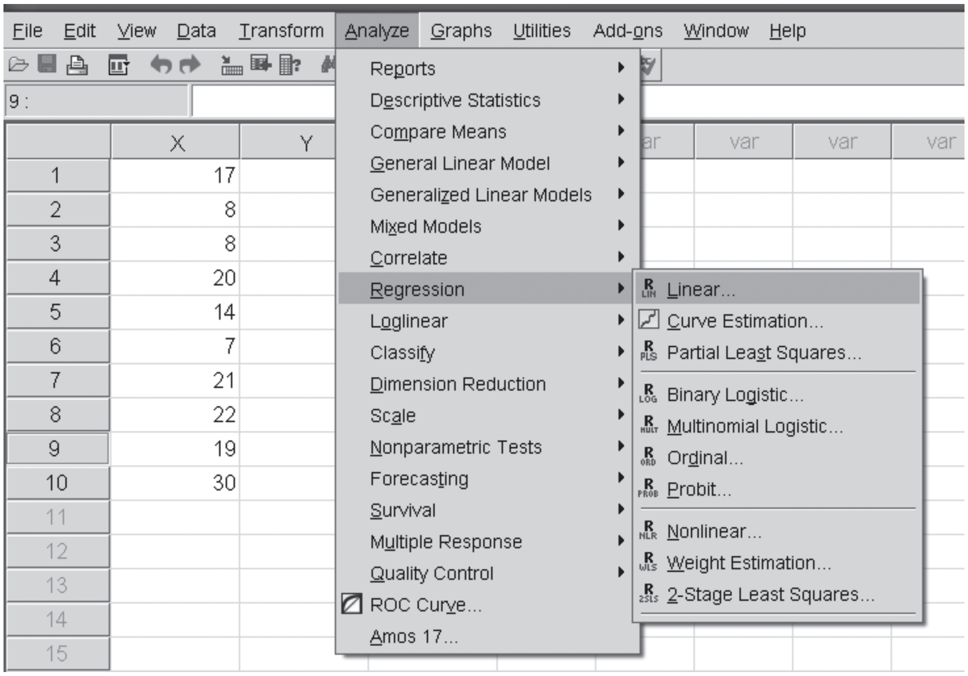
8.7 Calculating the "Raw-Score" Regression Equation and Standard Error of Estimate with IBM SPSS

Here, you will learn how to use IBM SPSS to calculate a regression equation and the standard error of estimate using the data in Tables 7.2 and 8.1. You will conduct a regression of *Y* on *X*. We will assume that you have already entered your data as described in Section 7.8.

To conduct a linear regression analysis in IBM SPSS, click **Analyze**, then select **Regression** followed by **Linear**. Figure 8.6 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze** → **Regression** → **Linear**.

Next, a dialog box will appear like that shown in Figure 8.7. You should move the Eating Difficulties variable (i.e., Y) to the box labeled Dependent. Then move the Stress variable (i.e., X) to the box labeled Independent(s). Figure 8.8 shows a screenshot after the *Y* and *X* variables have been correctly moved. Now click the OK button.

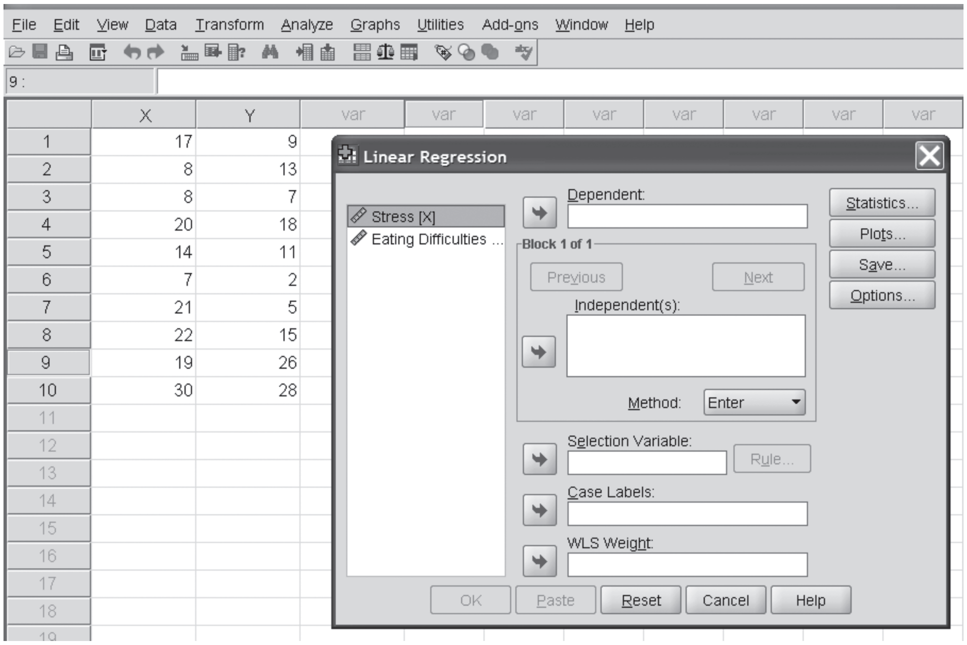
Table 8.2 shows the IBM SPSS output. You will notice that IBM SPSS produces a considerable amount of output by default. Three important sections are labeled Model Summary, ANOVA, and Coefficients. Because some of the output is beyond the scope of this section of the book, we will focus on selected aspects of the output. We will first focus on the bottom section of the output labeled Coefficients because it contains the regression equation for the best-fitting line. Under the column labeled B for Unstandardized Coefficients, the raw-score regression equation is $Y' = 0.783X + 0.4$. Under the column labeled Standardized Coefficients, the standard-score regression equation is $z'_Y = .675z_X$. Next, in the section of the output labeled Model Summary, IBM SPSS computed the standard error of estimate as 6.712. You will notice that it differs slightly from the 6.00 in Table 8.1. The


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
9 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
var
var
var
var
11
12
13
14
15

	X	Y
1	17	
2	8	
3	8	
4	20	
5	14	
6	7	
7	21	
8	22	
9	19	
10	30	

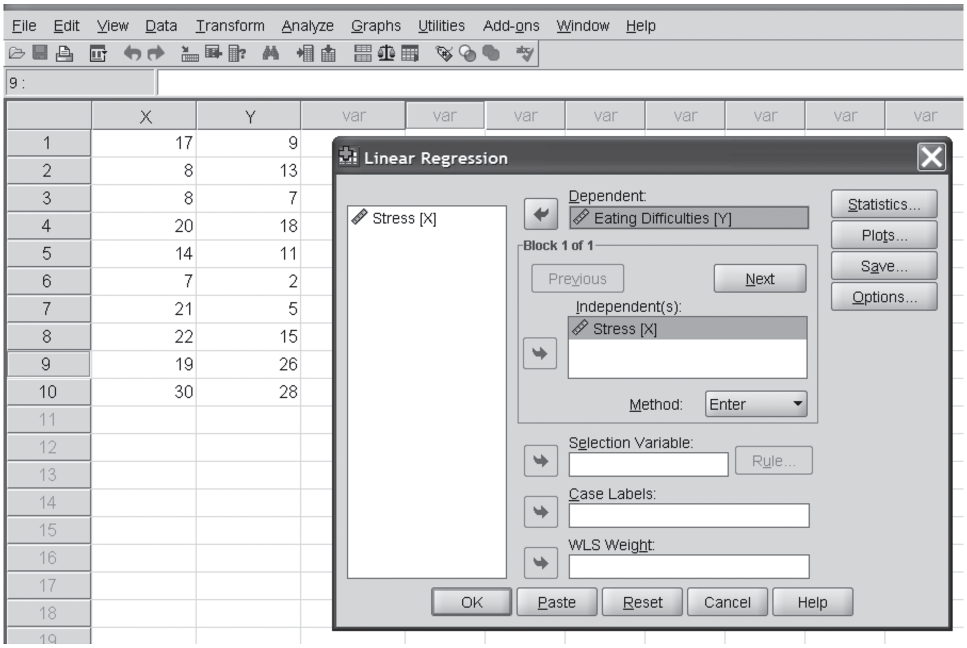
FIGURE 8.6 IBM SPSS menus when conducting a linear regression analysis.

reason for this is that we divided by n whereas IBM SPSS divides by a different value, $n - 2$ (see the analogous argument described in the Point of Controversy in Chapter 5 and Section 8.9). To get 6.00, square 6.712 and multiple by $(n-2)/n = 8/10$. Then, take the square root.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
9 :
var
Linear Regression
Stress [X]
Eating Difficulties ...
Dependent:
Block 1 of 1
Previous
Next
Independent(s):
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help
11
12
13
14
15
16
17
18
19

	X	Y
1	17	9
2	8	13
3	8	7
4	20	18
5	14	11
6	7	2
7	21	5
8	22	15
9	19	26
10	30	28

FIGURE 8.7 IBM SPSS dialog box from a linear regression analysis.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
9 :
var
Linear Regression
Stress [X]
Block 1 of 1
Previous
Next
Statistics...
Plots...
Save...
Options...
Selection Variable:
Rule...
Case Labels:
WLS Weight:
OK
Paste
Reset
Cancel
Help

	X	Y
1	17	9
2	8	13
3	8	7
4	20	18
5	14	11
6	7	2
7	21	5
8	22	15
9	19	26
10	30	28

Dependent: Eating Difficulties [Y]

Independent(s): Stress [X]

Method: Enter

FIGURE 8.8 IBM SPSS dialog box from a linear regression analysis after selecting variables.

8.8 Error in Estimating *Y* from *X*

If we can reasonably assume that the actual scores on Y are normally distributed about the predicted scores, we can use our knowledge of the normal curve to answer questions about errors of prediction. For example, using principles you learned in Chapter 6 and assuming a normal curve, we can determine that:

68% of the Y scores fall within the limits $\overline{Y} \pm 1.00S_Y$
95% of the Y scores fall within the limits $\overline{Y} \pm 1.96S_Y$
99% of the Y scores fall within the limits $\overline{Y} \pm 2.58S_Y$

Because a predicted score, Y', is a kind of mean and the standard error of estimate, S_{YX}, is a kind of standard deviation, it is also true, given a normal bivariate distribution, that:

68% of actual Y values fall within the limits $Y' \pm 1.00S_{YX}$
95% of actual Y values fall within the limits $Y' \pm 1.96S_{YX}$
99% of actual Y values fall within the limits $Y' \pm 2.58S_{YX}$

Suppose that the admissions office at a college applies this reasoning in predicting grades from SAT scores. Recall that our prediction for an applicant with an SAT total of 1,100 was 2.12. This prediction is only a best estimate of the mean of the distribution of GPAs for students whose SAT totals are 1,100. Some of those students will obtain higher GPAs, and some lower. To attach a margin of error to the prediction, the admissions office applies the principle that 95% of the values in a normal distribution fall within 1.96 standard deviations of the mean. Here, 2.12 serves as the mean, and the standard error of estimate serves as the standard deviation. The standard error of estimate, from Formula 8.4, is

$$S_{YX} = S_Y\sqrt{1 - r^2} = .50\sqrt{1 - .35^2} = .47$$

TABLE 8.2 *IBM SPSS Output from a Linear Regression Analysis*

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN (.05) POUT (.10)
  /NOORIGIN
  /DEPENDENT Y
  /METHOD=ENTER X.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.675[a]	.456	.388	6.712

[a] Predictors: (Constant), Stress

ANOVA[b]

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	301.965	1	301.965	6.702	0.32[a]
Residual	360.435	8	45.054		
Total	662.400	9			

[a] Predictors: (Constant), Stress
[b] Dependent Variable: Eating Difficulties

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
Model	B	Std. Error	Beta	t	Sig.
1 (Constant)	.400	5.452		.073	.943
Stress	.783	.302	.675	2.589	.032

[a] Dependent Variable: Eating Difficulties

Thus 95% of the actual GPAs will lie within these limits:

LOWER LIMIT: $Y' - 1.96S_{YX} = 2.12 - (1.96)(.47) = 1.20$

UPPER LIMIT: $Y' + 1.96S_{YX} = 2.12 + (1.96)(.47) = 3.04$

Look at Figure 8.9. It shows that for 95% of persons with a test score like our applicant's, freshman GPAs will fall between 1.20 and 3.04. If you are thinking that this margin of error is rather broad, you are not alone. Critics of the SAT argue that it does not sufficiently enhance the prediction of GPA beyond what rank in high school class allows (Slack & Porter, 1980).

The principles of Sections 6.4 and 6.5 enable us to answer other types of questions about prediction. For example, what proportion of students whose SAT total is 1,100 may we expect to achieve a GPA of 2.00 or below? To solve a problem of this kind, the first step is to find the predicted value of Y as we did in Section 8.4, where we found Y' to be 2.12.

We take the mean GPA (of students whose SAT total is 1,100) to be 2.12 and the standard deviation to be .47 (the value of S_{YX}). The problem now becomes one of the type described in Section 6.4. We must therefore translate the GPA of 2.00 to a z score and find

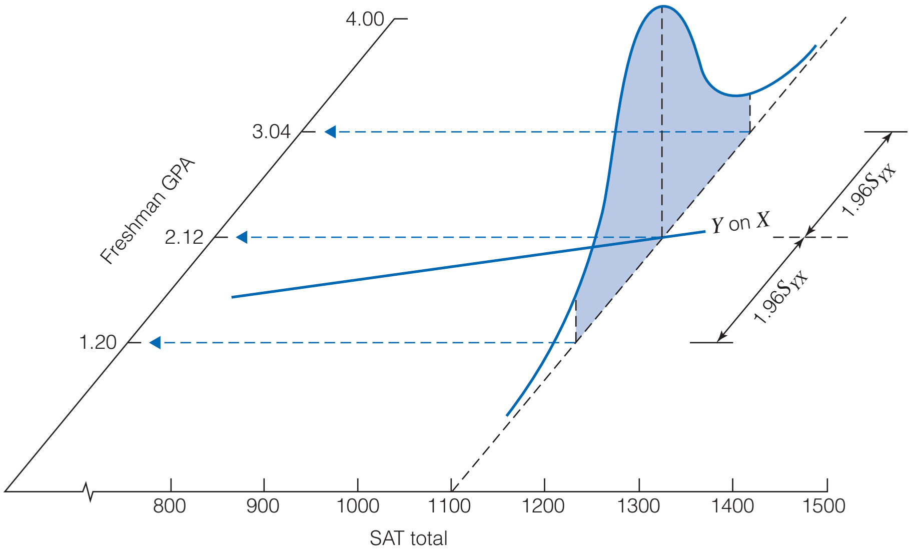


FIGURE 8.9 The 95% limits for actual GPAs of students scoring 1,100 on the SAT. The shaded area represents 95% of the cases with an SAT total of 1,100.

the area that lies below this point in a normal curve. Remembering that a z score is (*score − mean*)/*standard deviation*, we have:

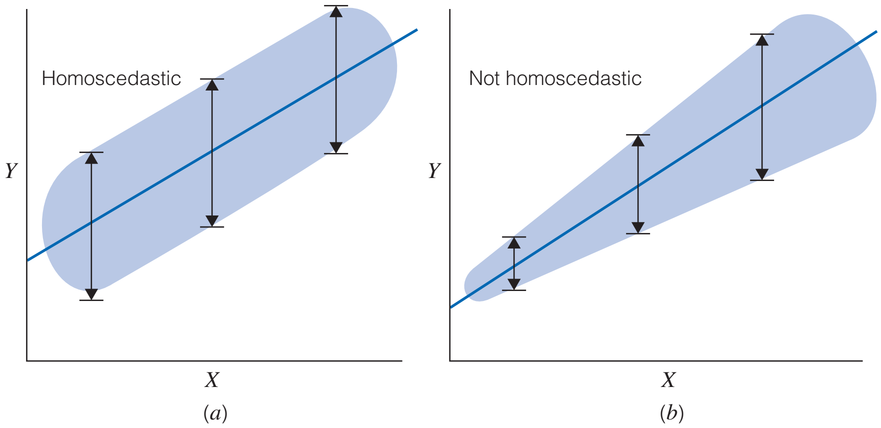
$$
\begin{aligned}
z &= \frac{Y - Y'}{S_{YX}} \\
&= \frac{2.00 - 2.12}{.47} \\
&= -.26
\end{aligned}
$$

Consulting Table A in Appendix D, we find that .3974 of the area under the curve falls beyond a z score of $-.26$ in a normal distribution. This is the answer to the problem: For those whose SAT total is 1,100, 39.74% may be expected to achieve a GPA of 2.00 or lower.

8.9 Cautions Concerning Estimation of Predictive Error

Several conditions must be met for the procedures of the preceding section to work well. First, *the relationship between the two variables must be linear.* Second, *the variability of actual* Y *values around* Y′ *must be the same for all values of* X. This is the assumption of **homoscedasticity**, the assumption that variability in Y is the same from column to column, as on the left side of Figure 8.10. In the defining formula for the standard error of estimate, $S_{YX} = \sqrt{(Y - Y')^2/n}$, we can see that S_{YX} is a function of the average of magnitude of the squared discrepancies between each Y value and the corresponding Y'. If the data are like those of Figure 8.7*b*, S_{YX} will approximate the standard deviation of the Y values only for intermediate values of X; it will overestimate predictive error in Y for low values of X and underestimate it for high values.

homoscedasticity the assumption that the variability of actual Y values around Y' is the same for all values of X



(*a*) (*b*)

FIGURE 8.10 (*a*) A homoscedastic bivariate distribution. The variability among the Y values is the same at all values of X, as suggested by the three double-headed arrows. (*b*) A bivariate distribution lacking homoscedasticity. The variability among the Y values differs at different values of X.

Third, *the actual* Y *scores must be normally distributed for all values of* X. Remember that for problems like those in the previous section, we used the table of areas under the normal curve to find the proportion of cases expected to fall above or below a given point.

The three conditions just listed must sound formidable. Fortunately, they are well met sufficiently often for the procedures of Section 8.8 to be reasonably accurate. We must now add one more warning: *The procedures of Section 8.8 do not take into account the influence of random sampling variation.* Because in any realistic prediction situation we are dealing not with the population of paired scores (why predict if we already know?) but with a sample from that population, the "true" correlation coefficient, regression line, and standard error of estimate will probably differ somewhat from those of the sample with which we are confronted. More sophisticated procedures are available that do allow for random sampling variation (see, for example, Hays, 1994).

8.10 Prediction Does Not Prove Causation

In Section 7.10, we discussed how correlation does not prove causation. As Figure 7.9 demonstrated, numerous relationships can underlie a correlation. Similarly, the fact that X can predict Y does not prove that X actually caused Y. We could have switched their roles, instead, predicting X from Y. We cannot distinguish the actual cause of an outcome simply based on estimating a regression equation and demonstrating that one variable can predict another. As an example, in industrial/organizational psychology, some research studies have demonstrated that job satisfaction (i.e., X) can predict job performance (i.e., Y). Other studies have found that job performance (i.e., X) can predict job satisfaction (i.e., Y). Some studies have suggested that job satisfaction and job performance may reciprocally cause one another. Yet other studies have noted that these two variables may simply be correlated and that they are both caused by other variables (see Judge, Thoresen, Bono, & Patton, 2001). Clearly, as this research suggests, computing correlation coefficients or regression equations (however well they fit) does not prove causation.

Notably, inferring causation can be a hotly debated topic. As highlighted by a number of statisticians and research methodologists, inferring causality is not a statistical issue (e.g., correlation or regression); it is an experimental design issue (Holland, 1986; Muthén, 1987; Rosopa & Stone-Romero, 2008; Tabachnick & Fidell, 2007).

8.11 Summary

If two variables are correlated, we can use the standing in one variable to predict the standing in the other. If the variables are linearly related, we can fit a straight line to the data and use that line (or its equation) to make such a prediction. Fortunately, linear relationships are frequently found in the real world.

The standard procedure is to establish the line according to the least-squares criterion, which specifies the line that minimizes the sum of squares of the predictive errors, $\Sigma(Y - Y')^2$. This line of best fit (called a regression line) estimates the mean value of Y for each particular value of X.

The regression equation can be expressed very simply in standard-score form: $z'_Y = rz_X$. This equation shows that (1) the regression line always passes through the point defined by the mean of X and the mean of Y no matter what the value of r; and (2) when the correlation is zero, the predicted value of Y is always the mean of Y, no matter what the value of X—an intuitively reasonable idea. The raw-score form of the regression equation and its corresponding graph is more useful for practical prediction work.

Unless the correlation is ±1, predicted values are only the expected values of Y corresponding to particular values of X, and we must consider the question of predictive error. The standard error of estimate, S_{YX}, measures that error. It is a kind of standard deviation, the standard deviation of the distribution of observed scores around the corresponding predicted score. Values of S_{YX} range from zero when the correlation is perfect to S_Y when there is no correlation at all. Our use of S_{YX} as a measure of predictive error in Y requires the assumption that the variability in Y is the same irrespective of the corresponding value of X. This is called the assumption of homoscedasticity.

If it is possible to assume that Y scores are normally distributed about their predicted scores, we can use the normal curve to answer questions about errors of prediction. Thus, we can estimate the limits that mark out the central 95% of Y scores that would be expected given a particular value of X.

Calculating a regression equation (or correlation) does not, by itself, prove causation. Establishing valid causal relationships is based on issues involving experimental design.

Mathematical Note

Note 8.1 Sum of the Deviations of Scores about Their Predicted Values (*Ref.*: Section 8.5)

Substituting the raw-score equivalent of Y' (from Formula 8.2) in the expression $\Sigma(Y - Y')$, we obtain

$$\sum(Y - Y') = \sum\left[Y - \left(r\frac{S_Y}{S_X}\right)X + \left(r\frac{S_Y}{S_X}\right)\overline{X} - \overline{Y}\right]$$

$$= \sum\left[(Y - \overline{Y}) - \left(r\frac{S_Y}{S_X}\right)(X - \overline{X})\right]$$

$$= \sum(Y - \overline{Y}) - r\frac{S_Y}{S_X}\sum(X - \overline{X})$$

Because $\Sigma(Y - \overline{Y}) = 0$ and $\Sigma(X - \overline{X}) = 0$ (see Note 4.1, Chapter 4), $\Sigma(Y - Y') = 0$.

Key Terms, Concepts, and Symbols

prediction (p. 125)
regression line (p. 126)
regression equation (p. 126)
assumption of linearity (p. 127)
Y' (p. 127)
d_Y (p. 127)
least-squares criterion (p. 127)
line of best fit (p. 127)
resistant line (p. 128)
regression line as a running mean (p. 128)
z'_Y (p. 129)
regression of Y on X (p. 129)
standard error of estimate (p. 132)
S_{YX} (p. 132)
homoscedasticity (p. 138)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Under what two circumstances is the prediction of Y from X better made from the regression equation than from the mean of the Y column corresponding to the particular value of X?

2. If $r = 1.00$, what is the value of Σd_Y^2? Why?

3. If $r = -1.00$, what is the value of Σd_Y^2? Why?

4. Suppose the correlation coefficient between the SAT verbal score and IQ is $r = +.50$. What z score on the SAT verbal test do you predict for someone whose z score on IQ is (a) $+3.00$? (b) -2.00? (c) 0.00? (d) -1.00? (e) $+1.00$? (f) -0.50?

5. Suppose $X = \text{IQ}$ with $\overline{X} = 100$ and $S_X = 15$, while $Y = \text{SAT verbal score}$ with $\overline{Y} = 500$ and $S_Y = 100$. Use the standard-score formula for the regression of Y on X to predict Y' (the raw score) for the following values of X, assuming $r = .50$. (You can do these in your head, but in writing your answer show both the z score and the raw-score values.) (a) 100, (b) 115, (c) 130, (d) 85, (e) 70, (f) 145, (g) 55.

6. Suppose $\overline{X} = 50$, $S_X = 10$, $\overline{Y} = 200$, and $S_Y = 30$. What is the value of r for each of the following circumstances? (*Hint*: You can do these mentally if you think in terms of the standard-score formula for the regression of Y on X.) (a) For $X = 40$, the regression equation predicts $Y' = 170$. (b) For $X = 40$, the equation predicts $Y' = 230$. (c) For $X = 60$, $Y' = 215$. (d) For $X = 30$, $Y' = 230$. (e) For $X = 70$, $Y' = 215$.

7. Suppose $\overline{X} = 55$, $S_X = 12$, $\overline{Y} = 118$, and $S_Y = 27$. For a case that stands at 55 on X, the regression equation predicts that $Y' = 118$. Can you tell the value of r? Explain.

8. According to the regression equation for a particular problem in prediction, when $X = 81$, $Y' = 317$; when $X = 84$, $Y' = 317$; and when $X = 92$, $Y' = 317$. (a) What is the value of r? How can you tell? (b) What is $\overline{Y}$? How can you tell? (c) Can you infer the value of $\overline{X}$? Explain.

9. If $r = 0$, then $Y' = \overline{Y}$, no matter what the value of X. What justification is there for such a prediction, aside from intuition?

10. For the example presented in Section 8.4, use the raw-score regression equation to predict the GPA for applicants with these SAT totals: (a) 900, (b) 800, (c) 1,250, (d) 1,400, (e) 1,600.

11. For the data in Problem 5, repeat parts (a), (c), and (e) using the raw-score regression equation.

12. Four subjects obtained the following scores on two measures:

Subject:	A	B	C	D
Score on X:	7	6	5	8
Score on Y:	9	6	7	3

(a) Find the raw-score regression equation, and predict the score on Y for these values of X: (b) 2, (c) 3, (d) 8, (e) 9.

13. For the data in Problem 4 of Chapter 7, (a) make a scatter diagram, (b) find the regression equation for raw scores, and (c) plot the regression line on your diagram. Now use the equation to predict the score on Y for these values of X: (d) 28, (e) 38, and (f) 47. (g) Verify that your predictions in (d) through (f) lie on the regression line in your diagram (show where they are on your graph).

14. For the 19 students in a statistics course, where X = height in inches and Y = weight in pounds, $\overline{X} = 69.18$, $S_X = 3.01$, $\overline{Y} = 147.42$, $S_Y = 24.61$, and $r = +.83$. (a) Write the raw-score formula for the regression of Y on X. Now predict the weight for students of these heights: (b) 60 in., (c) 62 in., (d) 68 in., (e) 72 in., (f) 80 in.

15. Using the data for Problem 14, (a) predict weight for a child who is 36 in. tall. (b) Any comments?

16. Calculate S_{YX} for the example of Section 8.4, using (a) Formula 8.3 and (b) Formula 8.4.

17. Calculate S_{YX} for the data in (a) Problem 5 and (b) Problem 14.

18. For the example of Section 8.4, find (a) the predicted GPA for an applicant whose SAT total is 1,300, (b) the limits within which 95% of the actual GPAs will probably lie for applicants with SAT totals of 1,300, and (c) the limits within which 68% of the actual GPAs will probably lie for applicants with SAT totals of 1,300.

19. The following data are for freshman students at Spartan University:

APTITUDE SCORE: X	FRESHMAN GPA: Y
$\overline{X} = 560$	$\overline{Y} = 2.65$
$S_X = 75$	$S_Y = .35$
$r_{XY} = +.50$	

(a) Write the raw-score regression equation for predicting Y from X. (b) John and Will score 485 and 710, respectively, on the aptitude test. Predict the freshman GPA for each. (c) What assumption is necessary for this prediction to be valid? (d) What is the value of the standard error of estimate of Y on X? (e) For students whose aptitude scores are the same as John's: What proportion will be expected to obtain a GPA equal to the freshman mean or better? (f) What proportion will be expected to obtain a GPA of 2.0 or lower? (g) Within what central GPA limits will 95% of students with aptitude scores equal to John's be likely to be found? For students whose aptitude score is the same as Will's: (h) What proportion will be expected to obtain a GPA equal to 2.5 or better? (i) What proportion will be expected to obtain a GPA of 3.0 or lower? (j) Within what central GPA limits will 50% of students with aptitude scores equal to Will's be likely to be found? (k) What assumptions are necessary for the answers to (e)–(j) to be valid?

20. Look at Figure 7.10. If we use the formulas for the regression equations given in this chapter, what will be the consequences of predicting Y (a) from an extremely low X score? (b) from an X score near the mean of the X values?

21. For the data in Problem 22 of Chapter 7, use IBM SPSS to calculate the (a) raw-score regression equation and (b) standard error of estimate. (Recall from Section 8.7 that the standard error of estimate will differ slightly between IBM SPSS and Formulas 8.3 and 8.4.)

22. For the data in Problem 22 of Chapter 7, use IBM SPSS to calculate (a) the standard-score regression equation. (Hint: In the Coefficients output of IBM SPSS, look for the column labeled Standardized Coefficients.) (b) In the output, why is the value for the constant empty?

23. Using the regression equation calculated in Problem 21, predict a participants score on Y (i.e., acts of aggression during the past year) if their score on X (i.e., mean number of media exposures per week) was: (a) 10, (b) 20, (c) 30, and (d) 40.

CHAPTER 9

Interpretive Aspects of Correlation and Regression

When you have finished studying this chapter, you should be able to:

- Understand the several different ways of interpreting the degree of correlation between two variables (e.g., the regression coefficient, coefficient of nondetermination, coefficient of determination, proportion of correct placements); and
- Explain what is meant by "regression toward the mean."

In the previous two chapters, we presented an introduction to the Pearson correlation coefficient and the regression equation, both based on the assumption of a linear relation between two variables. Along the way, we have warned that interpretation of the correlation coefficient is not always as straightforward as one might be tempted to believe. We know, for example, that a coefficient of .50 does *not* mean that there is 50% association between the two variables (Section 7.5). What, then, is the meaning of the correlation coefficient? Interpretation of Pearson's r and the regression equation is the subject of this chapter.

9.1 Factors Influencing r: Degree of Variability in Each Variable

The degree of variability characterizing each variable is an important consideration in understanding what a correlation coefficient means (or does not mean). First, let us lay a little groundwork.

According to Formula 8.4,

$S_{YX} = S_Y\sqrt{1 - r^2}$ the variability of *Y* scores around the regression line (sometimes called the error of prediction)

$$S_{YX} = S_Y\sqrt{1 - r^2}$$

We can solve this equation for r:

ALTERNATIVE FORMULA FOR r

$$r = \sqrt{1 - \frac{S^2_{YX}}{S^2_Y}} \qquad (9.1)$$

hugging principle the degree of correlation increases the more closely the *Y* scores "hug" the regression line

This is not a practical formula for computing the correlation coefficient, but it is helpful for understanding the meaning of r. Recall from Section 7.4 that if the Y scores "hug" the regression line closely, the correlation will be high, and if not, it will be low. We now give a more precise expression of this **hugging principle**. The magnitude of r is not a function of the size of S_{YX}, the absolute measure of the variation of Y about the regression line. Rather, it is a function of the size of S_{YX} relative to S_Y. If S_{YX} is zero, there is no error of prediction, and the correlation coefficient, according to Formula 9.1, is ± 1. On the other hand, if S_{YX} has the same value as S_Y, the correlation will be zero. Whether S_{YX}, the standard error of estimate, by itself is small or large therefore does not matter. What matters is whether it is small in relation to S_Y.

9.2 Interpretation of r: The Regression Equation I

Earlier (in Section 7.5) we observed that there was no simple way to interpret the meaning of the correlation coefficient. In this section, we consider the first of several ways to interpret r.

Recall from Section 8.3 that the equation for any straight line can be put into the following form:

GENERAL EQUATION
OF A STRAIGHT LINE

$$Y = bX + a \qquad (9.2)$$

***Y* intercept** the point on the *Y* axis at which a straight line crosses it

slope of a line specifies the amount of increase in *Y* that accompanies 1 unit of increase in *X*

In this equation, a is a constant that identifies the point on the Y axis at which the line crosses it. It is known as the ***Y* intercept**. For the two lines in Figure 9.1, the equation of one is $Y = 3X - 4$, and the equation of the other is $Y = .25X + 2$. Note that in each case the value of a identifies the point on the Y axis at which the line crosses it.

The other constant in the equation is b, and it is called the slope of the line. It will be positive when the value of Y increases with increasing values of X, as is the case with the two lines pictured. It will be negative if the line slants the other way—that is, when Y decreases as X increases. The **slope of a line** specifies the amount of increase in Y that accompanies one unit of increase in X. Consider the line for which the equation is $Y = 3X - 4$. The dashed lines show that if X is increased by 3 units, Y increases by 9 units. The ratio

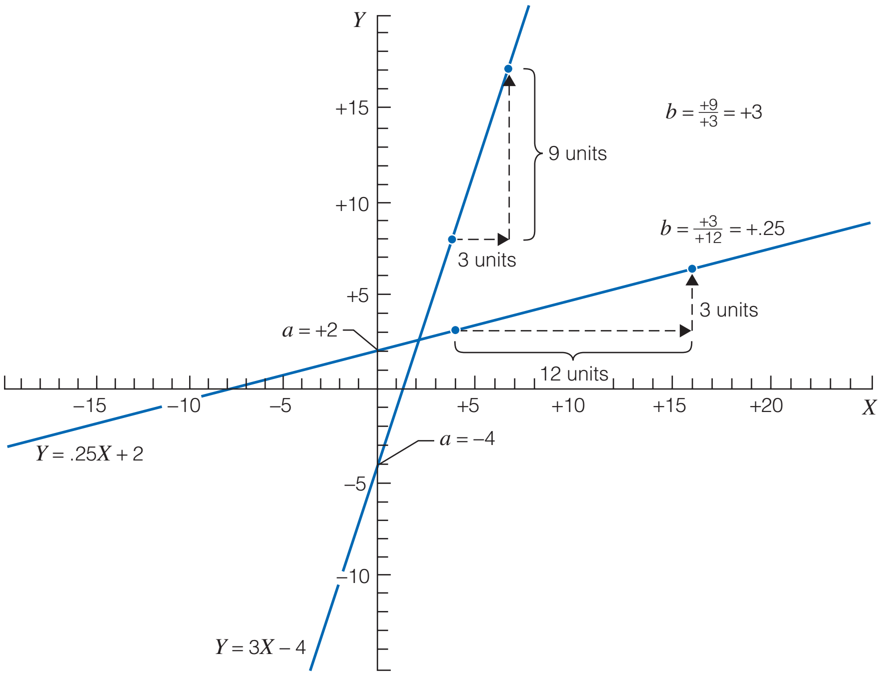


FIGURE 9.1 Plots of two straight lines showing the slope and *Y* intercept of each.

of the increase in *Y* (the vertical distance) to the increase in *X* (the horizontal distance) is the slope of the line. Some of you may have been told in your high school algebra class that the slope of a line was defined as "rise over run." Rise referred to the vertical change. Run referred to the horizontal change. Thus, the slope is

$$b = \frac{\textit{vertical change}}{\textit{horizontal change}} = \frac{+9}{+3} = +3.0$$

In the second equation, *Y* increases 3 units for every 12 units of increase in *X*, so the slope is $b = 3/12$ or $+.25$. Note that if *Y* decreased 3 units for every 12 units of increase in *X*, we would have a "negative increase" in the numerator of the fraction, and the resultant value of *b* would be $-.25$ rather than $+.25$.

Using this model, let us examine the several forms of the equation for the line of regression of *Y* on *X*:

GENERAL EQUATION OF A STRAIGHT LINE

$$Y = bX + a$$

REGRESSION OF *Y* ON *X*: STANDARD-SCORE FORMULA

$$z'_Y = \overbrace{rz_X}^{b} + \overbrace{0}^{(a\,=\,0)}$$

REGRESSION OF Y ON X: RAW-SCORE FORMULA

$$Y' = \overbrace{\left(r\,\frac{S_Y}{S_X}\right)}^{b} X + \overbrace{\left(-r\,\frac{S_Y}{S_X}\right)\overline{X} + \overline{Y}}^{a}$$

one interpretation of *r* the amount of increase in Y that, on average, accompanies a unit increase in X (when both measures are expressed in standard-score units)

regression coefficient the amount of increase in Y that, on the average, accompanies a unit increase in X (when both measures are expressed in raw-score units)

$r(S_Y/S_X)$ symbol for the regression coefficient

The value b is most interesting when the regression equation is cast in standard-score form. In this case, the correlation coefficient *is* the slope of the regression line. Therefore, *one interpretation of the correlation coefficient is that it states the amount of increase in* Y *that, on the average, accompanies a unit increase in* X *when both measures are expressed in standard-score units*. It indicates how much of a standard deviation Y will increase for an increase of 1 standard deviation of X.

When we state the regression equation for predicting Y from X in raw-score form, its slope is $b = r(S_Y/S_X)$ rather than r. This expression is known as the regression coefficient. It may also be used to interpret a given correlation. The **regression coefficient**, **$r(S_Y/S_X)$**, states the amount of increase in Y that, on average, accompanies a unit increase in X when both measures are expressed in raw-score units.

In his early work in measuring association, Sir Francis Galton apparently thought that the slope of the line of best fit might be an indicator of the degree of association. From what we have learned in this section, it is clear that this would work if the two standard deviations were equal because in that case, $r(S_Y/S_X)$ reduces to r. Indeed, much of Galton's work on heredity involved just such variables. Think back, for example, to Table 7.1, which contained the data from his investigation of the relationship between parental height and height of adult offspring. However, for many variables, the two standard deviations differ substantially (e.g., the association between GPA and SAT scores).

9.3 Interpretation of *r*: The Regression Equation II

In the previous section, we presented one way of using the regression equation to interpret r. Here is a second way. Consider the standard-score formula for the regression of Y on X: $z'_Y = rz_X$. Let us see what this equation says about the consequences of three possible values of r: 0, +.50, and +1.00. We show graphic representation of the regression of Y on X in Figure 9.2 for each of these situations together with the regression equation appropriate to each. As the diagram at the top shows, when the correlation is +1, the predicted standard score in Y is the same as the standard score in X from which the prediction is made. When the correlation is zero, the mean of $Y (z = 0)$ is predicted, no matter what value X has, as shown at the bottom. This is logical because information about X supplies no information that would improve on a sheer guess as to the value of Y.

The most interesting of the three situations is when r has an intermediate value. Think, for example, of the correlation between intelligence of parents and of offspring, which indeed has a value close to +.50. If parental intelligence is 2 standard deviations above the mean, the predicted intelligence of their offspring is 1 standard deviation above the mean. On the other hand, if the parents' intelligence is 2 standard deviations below the mean, the predicted intelligence of their offspring is only 1 standard deviation below the mean. To put it in other words, bright parents will tend to have children who are brighter than average but not as bright as they, and dull parents will tend to have children who are dull but not as dull as their

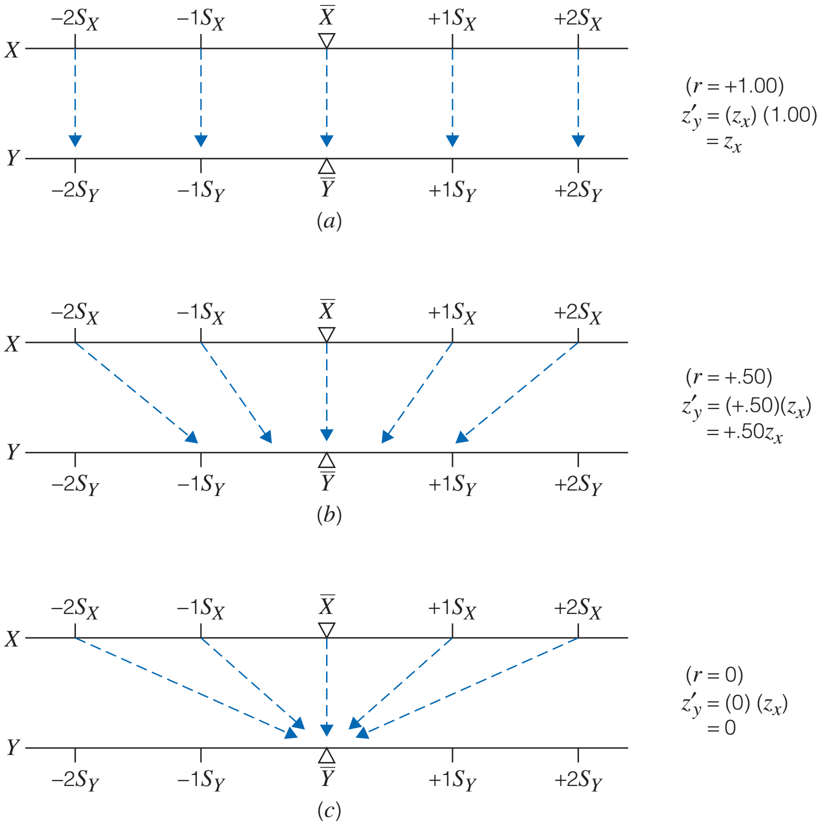


FIGURE 9.2 Regression toward the mean for three values of *r*.

parents. Remember that the predicted value is to be thought of as an *average* value; it is quite possible for bright parents to have a child brighter than they or one whose intelligence is below average.

This phenomenon is precisely what Galton observed in his studies of inheritance. He referred to it first as "reversion" and later as "regression." Today, we call it **regression toward the mean**. You understand now why the straight line of best fit is termed a *regression line* and why the symbol *r* was chosen for the correlation coefficient.

regression toward the mean when the correlation between two variables is less than perfect, the more extreme the score on one variable, the more likely it will be paired with a less extreme score on the other variable

Regression toward the mean is, of course, characteristic of *any* relationship in which the correlation is less than perfect. Consequently, we would expect that a very short person would tend to have short children but children less short than he or she, and that the brightest student in the class would tend to earn a good grade but not necessarily the best in the class.

Both the extremity of the value from which prediction is made and the value of *r* influence the amount of regression. First, *the more extreme the value from which prediction is made, the greater the amount of regression toward the mean*. In Figure 9.2*b*, the amount of regression is 1 standard deviation when prediction is made from a value 2 standard deviations above the mean, whereas regression is half that amount when prediction is made from a value 1 standard deviation above the mean.

Second, as the *z*-score form of the regression equation ($z'_Y = rz_X$) shows, *the higher the value of* r, *the less the amount of regression*. For $z_X = +1$, $z'_Y = +.9$ when $r = +.90$, $+.5$ when $r = +.50$, and $+.2$ when $r = +.20$.

If parents with extreme characteristics tend to have offspring with characteristics less extreme, how is it that, after a few generations, we do not find everybody at the center of the distribution? Galton's attention was attracted to this apparent paradox. The answer is that *the regression of predicted values toward the mean is accompanied by variation of obtained values about the predicted values, and the greater the degree of regression toward the mean, the greater the amount of variation*. The inward movement of regression is therefore accompanied by the expansive action of variability, which makes it possible for the more extreme values of Y to occur. Specifically, Y', the predicted value of Y, is only the predicted mean of Y for those who obtain a particular score in X. The obtained Y values corresponding to that value of X will be distributed about Y' with a standard deviation equal to S_{YX}. The lower the value of the correlation coefficient, the greater the value of S_{YX}. Thus, the greater the degree of regression toward the mean, the greater the variation of obtained Y values about their predicted values.

9.4 Interpretation of *r*: Proportion of Variation in *Y* Not Associated with Variation in *X*

Recall that the variability of Y scores about the regression line, or as we might call it, the error of prediction, is measured by the standard error of estimate, $S_{YX} = S_Y\sqrt{1 - r^2}$. The maximum possible error of prediction occurs when $r = 0$, in which case $S_{YX} = S_Y$. Therefore, one way we can interpret the correlation coefficient is by comparing the magnitude of error in the present predictive circumstances with that obtained in the worst possible predictive circumstances, that is, when $r = 0$. We have

$$\frac{\left[\begin{array}{c}\textit{magnitude of predictive error}\\ \textit{characterizing our predictive}\\ \textit{situation}\end{array}\right]}{\left[\begin{array}{c}\textit{magnitude of predictive error}\\ \textit{in the worst possible}\\ \textit{predictive situation (r = 0)}\end{array}\right]} = \frac{S_{YX}}{S_Y} = \frac{S_Y\sqrt{1 - r^2}}{S_Y} = \sqrt{1 - r^2} \qquad (9.3)$$

coefficient of nondetermination
$S_{YX}^2/S_Y^2 = 1 - r^2$; the proportion of the variance in Y that is *not* associated with differences in X

***k*²**
symbol for the coefficient of nondetermination

When this ratio is close to unity (its maximum value), the predictive error is close to its maximum. On the other hand, when the ratio is close to zero, most of the possible predictive error has been eliminated.

The ratio $S_{YX}^2/S_Y^2 = 1 - r^2$ shows the proportion of variation in Y that is not associated with differences in X (i.e., cannot be predicted from the regression equation). This is called the **coefficient of nondetermination** (symbolized ***k*²**). Older textbooks often referred to the ratio S_{YX}/S_Y (which equals $\sqrt{1 - r^2}$) as the coefficient of alienation (symbolized k). Today, it is more common to look at the variance in Y that *is* associated with differences in X (see the next section).

In Table 9.1 we present the coefficient of nondetermination for selected values of r. The table shows that the coefficient of nondetermination decreases very slowly as the correlation coefficient increases from zero. For instance, when the correlation coefficient has reached .50, the coefficient of nondetermination is still 75% of the size that it would be if the correlation were zero. Note also that a change in r from .20 to .30 results in a reduction in the coefficient of nondetermination of .05, whereas a change in r from .80 to .90 yields a reduction of .17. This shows that *a given change in the magnitude of a correlation coefficient has greater consequences when the correlation is high than when it is low.*

TABLE 9.1 *Several Interpretations of* r *for Different Levels of the Correlation Coefficient*

r_{XY}	k^2	r^2	A	B
1.00	.00	1.00	.50	1.00
.95	.10	.90	.40	.80
.90	.19	.81	.35	.71
.85	.28	.72	.32	.65
.80	.36	.64	.30	.59
.75	.44	.56	.27	.54
.70	.51	.49	.25	.49
.65	.58	.42	.23	.45
.60	.64	.36	.20	.41
.55	.70	.30	.19	.37
.50	.75	.25	.17	.33
.45	.80	.20	.15	.30
.40	.84	.16	.13	.26
.35	.88	.12	.11	.23
.30	.91	.09	.10	.19
.25	.94	.06	.08	.16
.20	.96	.04	.06	.13
.15	.98	.02	.05	.10
.10	.99	.01	.03	.06
.05	<1.00	>.00	.02	.03
.00	1.00	.00	.00	.00

k^2: coefficient of nondetermination (see Section 9.4)
r^2: coefficient of determination (see Section 9.5)
A: proportion of correct placements in excess of chance (see Section 9.6)
B: proportion of improvement in correct placement relative to the chance proportion of .50 (see Section 9.6)
Note: k^2, r^2, A, and B are unaffected by the sign (+ or −) of r.

The entries for Columns A and B of Table 9.1 are owed to W. B. Michael, "An Interpretation of the Coefficients of Predictive Validity and of Determination in Terms of the Proportions of Correct Inclusions or Exclusions in Cells of a Fourfold Table," *Educational and Psychological Measurement*, **26**, no. 2 (1966): 419–24.

9.5 Interpretation of r: Proportion of Variance in Y Associated with Variation in X

Assemble a number of scores; their values differ. What is the source of this variation? If we apply this question to the Y scores in a bivariate distribution, the answer has two parts. First, some of the variation in Y is associated with variation in X. That is, assuming some degree of correlation between X and Y, Y will likely have different values depending on whether the associated value of X is high or low. However, if we settle on any single value of X, there is still some variability in Y (this second kind of variability is measured by S_{YX}). *Total variation in* Y *may therefore be thought of as having two components: variation in* Y *that is associated with or attributable to differences in* X, *and variation in* Y *that is inherent in* Y *and, therefore, independent of differences in* X.

We can partition the total Y variation in such a way as to reflect the contributions of these two components. We must make the partition in terms of the *variance* of Y, which can be partitioned into additive components, rather than in terms of the standard deviation of

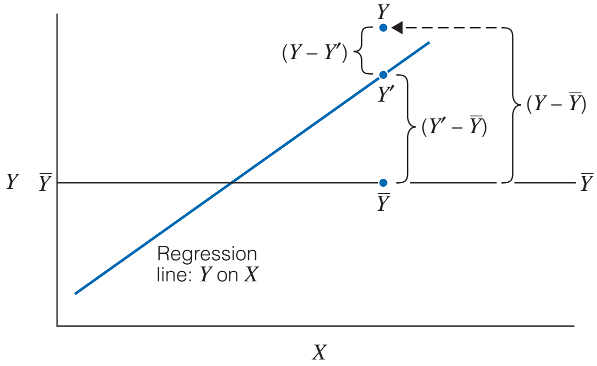


FIGURE 9.3 Partition of a deviation score $(Y - \overline{Y})$ into component deviations: $(Y - Y')$ and $(Y' - \overline{Y})$.

Y, which cannot. We begin by dividing the deviation of any Y score from its mean into two components:

$$(Y - \overline{Y}) = (Y - Y') + (Y' - \overline{Y})$$

Note that the equation is an algebraic identity because on the right side of the equation, the two values of Y' will cancel each other, if we wish, leaving the expression on the right the same as that on the left. Figure 9.3 illustrates the partition. We now have three deviation scores. Each measures variation of a particular kind:

$(Y - \overline{Y})$	This measures the total variation among Y scores. When these deviations are squared, summed, and divided by their number, we have: $S_Y^2 = \Sigma(Y - \overline{Y})^2/n$, the variance of the Y scores defined in the usual manner.
$(Y - Y')$	This measures the variation of obtained Y scores about the values predicted for them (Y'); it is the variation in Y that remains when X is held constant. Therefore, $(Y - Y')$ measures the Y variation that is independent of differences in X. When these deviations are squared, summed, and divided by their number, we have $S_{YX}^2 = \Sigma(Y - Y')^2/n$, the variance in Y that is independent of differences in X. This quantity is the square of the standard error of estimate (see Section 8.5).
$(Y' - \overline{Y})$	This quantity, unlike the other two, is new to us. It measures the variation of the predicted Y values about the mean of Y and, as explained shortly, measures the variation in Y that is associated with differences in X. When these deviations are squared, summed, and divided by their number, we have: $S_{Y'}^2 = \Sigma(Y' - \overline{Y})^2/n$, the variance in Y that is associated with differences in X.

A few more words about $S_{Y'}^2$ are in order. First, look at Figure 9.3 and find the discrepancy $(Y' - \overline{Y})$ from which this variance is generated. What would happen if $r = 0$? In this case, the predicted value of Y is $\overline{Y}$, irrespective of the value of X from which the prediction is made (see Section 8.3). Then all values of $(Y' - \overline{Y})$ would become $(\overline{Y} - \overline{Y})$, and $S_{Y'}^2$ becomes zero. In other words, when $r = 0$, none of the Y variance is associated with variation in X. On the other hand, if $r = 1.00$, each predicted value of Y is the actual value of Y. Then, each value of $(Y' - \overline{Y})$ would become $(Y - \overline{Y})$, and $S_{Y'}^2$ becomes S_Y^2. This would mean that *all* the variation in Y is associated with variation in X. We can express the relationships developed above in the following equation:

PARTITION OF *Y* VARIANCE
INTO TWO COMPONENTS

$$S_Y^2 = S_{YX}^2 + S_{Y'}^2 \qquad (9.4)$$

$$\begin{bmatrix}\textit{total}\\ \text{Y}\\ \textit{variance}\end{bmatrix} = \begin{bmatrix}\textit{variance in } \text{Y}\\ \textit{independent of}\\ \textit{variation in } \text{X}\end{bmatrix} + \begin{bmatrix}\textit{variance in } \text{Y}\\ \textit{associated with}\\ \textit{variation in } \text{X}\end{bmatrix}$$

partition of *Y* variance (S_Y^2) into two components
1. variance in *Y* independent of variation in *X* ($=S_{YX}^2$), and 2. variance in *Y* associated with variation in *X* ($=S_{Y'}^2$)

Another way we can interpret the correlation coefficient is in terms of the *proportion of the total* Y *variance that is associated with differences in* X. This proportion is given by:

$$\frac{[\textit{variance in } \text{Y} \textit{ associated with differences in } \text{X}]}{[\textit{total variance in } \text{Y}]} = \frac{S_{Y'}^2}{S_Y^2}$$

This expression can be simplified by appropriate substitution. Because

$$S_Y^2 = S_{YX}^2 + S_{Y'}^2$$

then

$$S_{Y'}^2 = S_Y^2 - S_{YX}^2$$

and

$$\frac{S_{Y'}^2}{S_Y^2} = \frac{S_Y^2 - S_{YX}^2}{S_Y^2}$$

$$= \frac{S_Y^2 - \left(S_Y\sqrt{1-r^2}\right)^2}{S_Y^2}$$

$$= \frac{S_Y^2 - S_Y^2(1-r^2)}{S_Y^2}$$

$$= \frac{S_Y^2 - S_Y^2 + S_Y^2 r^2}{S_Y^2}$$

$$= r^2$$

r^2
symbol for the coefficient of determination

coefficient of determination
$S_{Y'}^2/S_Y^2 = r^2$; the proportion of the *Y* variance that is associated with differences in *X*

Thus, **r^2**, called the **coefficient of determination**, gives the proportion of the *Y* variance that is associated with differences in *X*. If $r = +.50$, for example, $r^2 = .25$. This means that 25% of the *Y* variance is associated with differences in *X* (and 75% is not). Values of r^2 are given for selected values of *r* in Table 9.1. Notice, once more, that the proportion of *Y* variance accounted for by variation in *X* increases more slowly than does the magnitude of the correlation coefficient. Not until $r = +.71$ does $r^2 = .50$.

Notice that the manner of interpreting *r* in this section is closely related to the manner we presented in the last section ($k^2 = 1 - r^2$). Thus,

$$\frac{S_Y^2}{S_Y^2} = k^2 + r^2 \qquad (9.5)$$

$$\begin{bmatrix}\textit{total } \text{Y} \textit{ variance}\\ \textit{expressed as a}\\ \textit{proportion}\end{bmatrix} = \begin{bmatrix}\textit{proportion of } \text{Y}\\ \textit{variance independent}\\ \textit{of } \text{X} \textit{ variance}\end{bmatrix} + \begin{bmatrix}\textit{proportion of } \text{Y}\\ \textit{variance associated}\\ \textit{with } \text{X} \textit{ variance}\end{bmatrix}$$

Look again at Table 9.1 to see the relationship between k^2 and r^2.

TABLE 9.2 *Proportion of Successful Placements for Two Levels of* r

(r = .00)

Y: Criterion		BELOW X MEDIAN	ABOVE X MEDIAN
	ABOVE Y MEDIAN	II 50	I **50**
	BELOW Y MEDIAN	III **50**	IV 50

$n = 200$

X: PREDICTOR

(r = +.50)

Y: Criterion		BELOW X MEDIAN	ABOVE X MEDIAN
	ABOVE Y MEDIAN	II 33	I **67**
	BELOW Y MEDIAN	III **67**	IV 33

$n = 200$

X: PREDICTOR

9.6 Interpretation of r: Proportion of Correct Placements

The interpretation of r according to k^2 or r^2 is certainly not very encouraging as to the use of values of r of moderate magnitude. However, in practical problems of prediction, we are often interested in estimating, from a person's score in X, the likelihood that she or he will "succeed" (i.e., score above a given point in Y). In these cases, we can often do much better by using the predictor than by not doing so, even when the correlation is modest. We introduced this manner of interpreting r in Section 7.5 and expand on it here.

Assume a normal bivariate distribution of 200 applicants, that success is defined as scoring above the median on the criterion variable (Y), and that those who are selected as potentially successful are those who score above the median on the predictor variable (X). If $r = 0$, 50% correct placement will be achieved by using the score on the predictor for selection. Look at the first picture in Table 9.2. Correct predictions have been made for the 50 individuals who were predicted to succeed and did, plus the 50 who were predicted to fail and did, that is, those individuals who fall, respectively, in Quadrants I and III of the diagram. The proportion of correct placements is therefore 100 out of the 200 candidates or 50%.

Now look at the second picture in Table 9.2. In this case, the correlation coefficient is +.50, and the number of correct placements is 67 + 67 = 134. The proportion of correct placements is, therefore, 134/200 or .67. The proportion of correct placements in excess of chance (.50) is .17. The proportion of improvement in correct placement relative to chance is .17/.50 or .33 (.17/.50 actually equals .34, but .33 is the right value; the discrepancy is due to rounding). Viewed this way, a correlation coefficient of .50 appears to have greater use than the interpretations of Sections 9.4–9.5 suggest.

The last two columns of Table 9.1 present, for selected values of r, the proportion of correct placements in excess of chance and the proportion of improvement in correct placements relative to the chance value of .50. This interpretive approach and the entries for A and B of Table 9.1 are owed to Michael (1966).

9.7 Summary

The facts that the correlation coefficient is zero when two variables are linearly unrelated and ±1 when the linear relation is perfect suggest that interpreting the degree of association must be an easy task. Alas, we have found it to be much more complex than we might have supposed.

Although the absolute value of the correlation coefficient varies from zero to one, the degree of association is, as stated earlier, not ordinarily interpretable in direct proportion to its magnitude. Nevertheless, we have found several ways of saying what a given coefficient means. The slope of the straight line of best fit gives us one way: it specifies the amount of increase (or decrease) in one variable that accompanies 1 unit of increase in the other variable (Section 9.2). A second interpretation may be made through the ratio S^2_{YX}/S^2_Y (called k^2, the coefficient of nondetermination) (Section 9.4). This gives information about the reduction in the magnitude of the predictive error accomplished by using the knowledge of the extent of association rather than by just guessing. A third interpretation is given by examining r^2, the coefficient of determination (Section 9.5). It states the proportion of variance of the one variable that is associated with differences in the other. Study of the behavior of both k^2 and r^2 shows that a change of a given magnitude (.10, for example) in the correlation coefficient has greater consequence when it is applied to coefficients having high values than to those having low values. Finally, another way of interpreting correlation (and prediction) is in terms of the proportion of correct placements expected when the regression equation is used in practical problems of prediction (Sections 7.5 and 9.6).

These several ways of interpreting correlation and prediction are quite diverse. It is clear not only that the problem of interpretation is not simple but also that there is no one "best" way. The particular nature of the problem, theoretical or applied, and the context in which it is set must guide the investigator to an appropriate choice.

Key Terms, Concepts, and Symbols

hugging principle (p. 144)

Y intercept (p. 144)

slope of a line (p. 144)

regression coefficient (p. 146)

$r(S_Y/S_X)$ (p. 146)

regression toward the mean (p. 147)

coefficient of nondetermination (p. 148)

k^2 (p. 148)

partition of variance (p. 150)

total Y variance (p. 150)

$(Y - \overline{Y})$, $(Y - Y')$, $(Y' - \overline{Y})$ (p. 150)

variance in Y independent of variation in X (p. 151)

variance in Y associated with variation in X (p. 151)

S^2_Y, S^2_{YX}, $S^2_{Y'}$ (p. 151)

r^2 (p. 151)

coefficient of determination (p. 151)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Solve the equation $S_{YX} = S_Y\sqrt{1 - r^2}$ for r, thus obtaining Formula 9.1.

2. One regression equation reads: $Y' = 2X - 5$. (a) Could the X and Y scores to which this equation refers be in standard-score form? Explain. (b) Interpret r in terms of the slope of the regression equation.

3. Another regression equation reads: $Y' = -.25X + 5$. (a) Could the X and Y scores to which this equation refers be in standard-score form? Explain. (b) Could r be negative? Explain. (c) Is $r = -.25$? Explain. (d) Interpret r in terms of the slope of the regression equation.

4. Another regression equation reads: $Y' = 3X$. Is that possible? Explain.

5. Another regression equation reads: $Y' = 12$. State all the things that we know from this about X, Y, and/or r.

6. Section 8.4 presented an equation for predicting GPA (Y) from SAT total score (X): $Y' = .00125X + .75$. (a) What is the regression coefficient for this equation? (b) Interpret the regression coefficient in terms of the slope of the regression line.

7. For the data used as an example in Section 8.4, the equation for predicting SAT total score (X) from GPA (Y) is $X' = 98Y + 804$. (a) What is the regression coefficient for this equation? (b) Interpret the regression coefficient in terms of the slope of the regression line.

8. For the data from Chapter 8 used in Problem 14, (a) write the raw-score formula for the regression of weight on height. (b) What is the regression coefficient for this equation? (c) Interpret the regression coefficient in terms of the slope of the regression line.

9. Someone proposes a "hothouse" treatment for gifted children. Four-year-olds are tested, and those whose IQs appear to be above 140 are selected for special treatment. Before treatment, mean IQ of this group is 150. Let us suppose that treatment has no effect on IQ. What mean IQ would you expect when the children are retested 1 year later? Explain.

10. In one company, it is found that the frequency of accidents correlates about $+.40$ from one year to the next. The company institutes a safety course for the 10% least safe workers last year. This year, the accident record of these people is distinctly better. Your comment?

11. The correlation between sociability and impulsiveness is about $+.6$ among college students. (a) In a large sample of college students, do you expect those in the top 5% on sociability all to be in the top 5% on impulsiveness? Explain. (b) Do you expect those in the bottom 5% on sociability to be in the bottom 5% on impulsiveness? Explain.

12. For the relationship between SAT score and freshman-year GPA (see Section 8.4), (a) what is the coefficient of nondetermination? (b) Explain what this value means. (c) What is the coefficient of determination? (d) Explain what this value means.

13. For the data presented in Tables 7.2 and 7.4, (a) what is the coefficient of nondetermination? (b) Explain what this value means. (c) What is the coefficient of determination? (d) Explain what this value means.

14. When $r = \pm 1.00$, what value would $\Sigma(Y - Y')^2$ have? Explain.

15. When $r = 0$, what value would $\Sigma(Y - Y')^2$ have? Explain.

16. Under what conditions will $S_Y = S_{Y'}$? Explain.

17. Under what conditions will $S_Y = S_{YX}$? Explain.

18. Given the data in Section 8.4 on SAT scores and college GPAs, (a) use Table 9.1 to determine how well an admissions office can expect to do in predicting which of their applicants would earn a GPA in the top half of the freshman class. (b) How well would they do if they flipped a coin to make the prediction (heads predicts top half, tails predicts bottom half)? (c) What assumptions underlie the answer to (a)?

19. Interpret the correlation coefficients $+.10$, $+.30$, $+.60$, and $+.90$ in terms of (a) the coefficient of nondetermination, (b) the coefficient of determination, and (c) the proportion of correct placements when the split is at the median in X and Y.

CHAPTER 10

Probability

When you have finished studying this chapter, you should be able to:

- Understand that the probability of an event is the relative frequency with which that event occurs over an infinite number of repetitions of the process that generated it;
- Know that as a proportion, the probability of an event lies between 0 (an impossible event) and 1 (a certain event);
- Know that probability does not tell us what will happen in the short run;
- Estimate the probability of event A by assuming equally-likely outcomes and dividing the number of possible outcomes of A by the total number of possible outcomes;
- Explain the addition theorem and the multiplication theorem of probability; and
- Understand what is meant by a probability (theoretical relative-frequency) distribution, and explain why the binomial distribution is an example of a probability distribution.

Have you ever bought a lottery ticket and dreamed of what you would buy with all that money if you actually won? We all know, of course, that because millions of lottery tickets are usually purchased before each drawing, our chances of winning are very remote. But for a few people—a very few—the dream really does come true. It came true for Evelyn Marie Adams in 1986. She entered and won the New Jersey Lottery, making her an instant millionaire. State officials said that for a person buying a single ticket, the probability of winning was about 1 in 4.123 million. What makes the story interesting is that 4 months later, Ms. Adams purchased another ticket in the New Jersey Lottery and won again. Statisticians quickly calculated that the probability of Ms. Adams winning both times (assuming that she had bought a single ticket each time) was 1 in 17 trillion!

Is this an amazing, unbelievable coincidence? For Evelyn Marie Adams it was a 1-in-17-trillion coincidence, but statisticians Perci Diaconis and Frederick Mosteller of Harvard University (1989) claimed that the probability of someone somewhere in the United States winning a lottery twice within 4 months was really closer to 1 in 30. According to them, Ms. Adams' accomplishment was not quite that amazing after all.

Which is correct: 1 in 17 trillion or 1 in 30? Actually, both are correct. To fully understand why, let us first develop the concept of probability.

10.1 Defining Probability

What is the probability that when:

A coin is tossed, it will come up heads?

A die is rolled, it will result in a 4?

A card is selected blindly from a deck of 52 cards, it will be an ace?

event
an observable outcome

generating event
any repeatable process that results in only one outcome at a time

probability (Pr) of an event
the relative frequency with which that event would be observed over an *infinite* number of repetitions of the generating event, when each generating event is conducted in the same manner

To answer questions like these, let us begin with a few definitions. An **event** is an observable outcome. A **generating event** is any repeatable process that results in only *one* outcome at a time. For example, tossing a coin is a generating event, and the outcome (heads or tails) is an event.

In each of the preceding examples, it is possible to think of a generating event (e.g., a toss of a particular coin) as repeating indefinitely. One way of viewing the probability of an event is as the proportion of time that the event occurs in the long run. Here, the "long run" means an infinite number of trials. Thus, the **probability (Pr) of an event** is the relative frequency with which that event would be observed over an infinite number of repetitions of the generating event, when each generating event is conducted in the same manner. By this definition, the probability of observing heads when we toss a particular coin is

$$\text{Pr (heads)} = \frac{\textit{number of observations of heads}}{\textit{total number of observations}} = \frac{f\ (\textit{of heads})}{N}$$

when observations are recorded for an infinite number of tosses. As a proportion, probability lies between 0 (an impossible event) and 1 (a certain event).

Because a probability is a proportion characterizing an infinitely long series of occurrences, it does not tell us precisely what will happen over any short run. Even if it is true that the probability of heads is 1/2 for a toss of a particular coin, that does *not* mean that in the next 2 tosses, 1 of the outcomes must be heads, or that in 100 tosses, 50 will be heads. However, if we repeat the generating event over and over, the relative frequency of observations of heads would approach 1/2 (but would probably not be *exactly* equal to 1/2 even after 1 million tosses) and would be exactly 1/2 for an infinite number of tosses.

Now let us return to our initial question: What is the probability that when we toss a coin it will come up heads? The probability does not have to be 1/2 just because there are two possible outcomes. Maybe the coin is bent, resulting in heads coming up only 45% of the time in an infinitely long series of tosses. If tossing a certain coin results in heads on 45% of the tosses in the long run, then the probability of the event heads is .45 for a toss of that coin.

"HOW DO YOU WANT IT—THE CRYSTAL BALL MUMBO-JUMBO OR STATISTICAL PROBABILITY?"
S. Harris

© Sidney Harris, ScienceCartoonsPlus.com

empirical probability an estimate of the probability of an event, obtained by finding the frequency of the event over a large number of repetitions of the generating event

So how do we know the probability of obtaining heads on the toss of a particular coin? To determine the precise probability, we would have to toss the coin an infinite number of times. This, of course, is not humanly possible. We might toss a coin several thousand times to obtain an estimate of the true value, called an **empirical probability**, but in inferential statistics we will be more concerned with outcomes of experiments than with outcomes of coin tosses. No researcher would wish to repeat an experiment thousands of times to closely estimate the probability of an outcome. Fortunately, there is a mathematical model that we can use to determine long-run probabilities without having to repeat the generating event an infinite number of times. However, using this model requires that we make an important assumption, which we discuss next.

10.2 A Mathematical Model of Probability

equally-likely model of probability given a population of possible outcomes, each of which is equally likely to occur, Pr (A) =

$$\frac{\text{number of possible A outcomes}}{\text{total number of possible outcomes}}$$

If we can assume that the primary generating event yields equally-likely outcomes, we can solve for the probabilities of more complex events without actually having to make an infinite number of observations. We will call this way of looking at probability the **equally-likely model of probability**:

> Given a population of possible outcomes, each of which is equally likely to occur, the probability of event A on a single trial is equal to the number of possible outcomes yielding A, divided by the total number of possible outcomes.

In a deck of 52 cards, there are four suits of 13 cards each. Spades and clubs are black, and diamonds and hearts are red. If we draw cards from a deck (the generating event) in such a manner that each card is equally likely to be selected, what is the probability of drawing a spade?

$$\text{Pr (spade)} = \frac{\textit{number of possible spade outcomes}}{\textit{total number of possible outcomes}} = \frac{13}{52} = \frac{1}{4}$$

Similarly, the probability of drawing an ace is 4/52, the probability of drawing an ace in a red suit is 2/52, and the probability of drawing a card (of any denomination) in a red suit is $26/52 = 1/2$.

If we assume that the two outcomes of tossing a coin (heads or tails, no edge landings) are equally likely, then

$$\text{Pr (heads)} = \frac{\textit{number of possible heads outcomes}}{\textit{total number of possible outcomes}} = \frac{1}{2}$$

Compare this ratio for Pr (heads) with that given in the previous section. If the assumption that the generating event yields equally-likely outcomes is true, this ratio gives the same value as that obtained for relative frequency. When the assumption is not true, the probability values provided by the equally-likely model will not be accurate. For example, a bent coin may have the long-run probability of heads of .45, but the equally-likely model will yield a value of .5. Fortunately, many real events are close enough to the assumption of equally-likely outcomes that we can use the equally-likely model to determine probabilities.

Most importantly, these approaches are very useful as preliminary guides for considering how to make decisions in the face of uncertainty—an important part of inferential statistics.

10.3 Two Theorems in Probability

addition theorem of probability (the "or" rule) the probability of occurrence of any one of several particular events is the sum of their individual probabilities (provided that they are mutually exclusive)

Let us look again at a problem introduced in the previous section. In a deck of 52 cards, what is the probability of drawing an ace? We observed that there are four ways of drawing an ace out of the 52 ways in which a card might be drawn. According to the principle of relative frequency among equally-likely outcomes, the probability of drawing an ace is 4/52.

The same problem may be viewed another way. The probability of drawing an ace of spades is 1/52, and the same is true for the ace of hearts, the ace of diamonds, and the ace of clubs. The probability of drawing the ace of spades *or* the ace of hearts *or* the ace of diamonds *or* the ace of clubs is the sum of the individual probabilities, that is, $1/52 + 1/52 + 1/52 + 1/52 = 4/52$. Similarly, the probability of obtaining a 5 *or* a 6 on a single toss of a die is $1/6 + 1/6 = 2/6$. These examples illustrate the **addition theorem of probability**:

> The probability of occurrence of any one of several particular events is the sum of their individual probabilities, provided that they are mutually exclusive.

mutually exclusive the occurrence of one event precludes the occurrence of any others

You can usually recognize when to apply the rule by the use of the word *or* (some refer to the addition theorem as the "*or*" *rule*). Notice that the rule is valid only when the outcomes are **mutually exclusive**; that is, when the occurrence of one precludes the occurrence of any of the others. In the examples above, the card drawn may be the ace of spades *or* the ace of hearts, but it cannot be both. Similarly, a die may come up 5 *or* 6, but we cannot have it both ways. Note 10.1 at the end of the chapter presents a modification of the addition theorem that is appropriate when outcomes are not mutually exclusive.

Another theorem concerns successive and joint events. If we toss the same coin twice, what is the probability of heads coming up on the first toss *and* again on the second toss? Let us view the tossing of the same coin twice as an experiment that is to be repeated indefinitely. In the long run, the first toss will come up heads half the time. The second toss

multiplication theorem of probability (the "and" rule)
the probability of several particular events occurring successively or jointly is the product of their separate probabilities (provided that the generating events are independent)

independent events
the outcome of one event has no influence on and is not related to the other events

will come up heads in half of those trials in which the first toss came up heads. Therefore, the proportion of *all* trials in which both tosses will come up heads is (1/2)(1/2) = 1/4. The probability of obtaining two heads in a row is thus 1/4. The same is true of joint (simultaneous) events: The probability of obtaining two heads if we toss two coins at once is 1/4. This illustrates the **multiplication theorem of probability** (which some call the "*and*" *rule*):

> The probability of several particular events occurring successively or jointly is the product of their separate probabilities, provided that the generating events are independent.

Again, we have a qualification: The generating events must be **independent events**. This means that the outcome of one event must have no influence on and must in no way be related to the outcome of the other event. (Formal consideration of independence in a probability sense is given in Note 10.3, but read Note 10.2 first.) For coin tosses, the way the first toss falls has no influence on the way the second toss will fall. The outcome of the two tosses is therefore the outcome of two independent events. Notice that the rule may be extended to more than two independent events. If we ask the probability of obtaining three heads on three successive tosses of a coin, it is $(1/2)(1/2)(1/2) = (1/2)^3 = 1/8$.

People being what they are, it is not unusual for them to act as though events are not independent when in fact they are. This failure occurs so often that it has been given a name: the *gambler's fallacy*. If we look over the shoulder of a person playing roulette, a game in which one guesses which of 38 numbers a ball will land on, we may find him keeping track of the numbers that have come up, with the idea that he will bet on one that has not appeared recently and so is "due." Such a system is useless. Assuming a fair wheel, fairly spun, there is no way in which the outcome of an earlier trial may affect the outcome of the present trial. The wheel has no memory.

Those just getting acquainted with probability sometimes become confused about mutually exclusive outcomes and independent events. Look carefully at the explanations that accompany their introduction; they do not mean the same thing. The concept of independence applies to a pair of generating events (e.g., two coin tosses), not to specific outcomes. For example, the particular outcomes that may result from a *single* trial are always completely dependent if they are mutually exclusive: The occurrence of one predicts with certainty that the other(s) will *not* occur. The multiplication theorem applies only when two or more events are considered together, as in the tossing of two coins or the tossing of one coin twice.

10.4 An Example of a Probability Distribution: The Binomial

If we toss a coin two times, what are the possible outcomes? Assuming that heads or tails are the only possible outcomes, we could get two heads in succession, two tails in succession, a head followed by a tail, or a tail followed by a head (HT and TH are two *different* outcomes). We find the probability of any of these successive outcomes by the multiplication theorem: (1/2)(1/2) = 1/4. The addition theorem gives the probability of obtaining one head and one tail in either order (HT or TH): 1/4 + 1/4 = 1/2. The resultant probabilities are shown in Figure 10.1*a*.

Suppose we toss the coin three times. How many possible outcomes are there now? There are eight: three tails in succession, three combinations of two tails and one head (TTH, THT, and HTT), three combinations of two heads and one tail (HHT, HTH, THH), and three heads in succession. Which outcome is more likely, HHH or HTH? Think about it a moment. The probability is the same for each: (1/2)(1/2)(1/2) = 1/8. Which outcome is more likely, HHH or two heads and a tail in any order? This is a different question, for the latter includes the

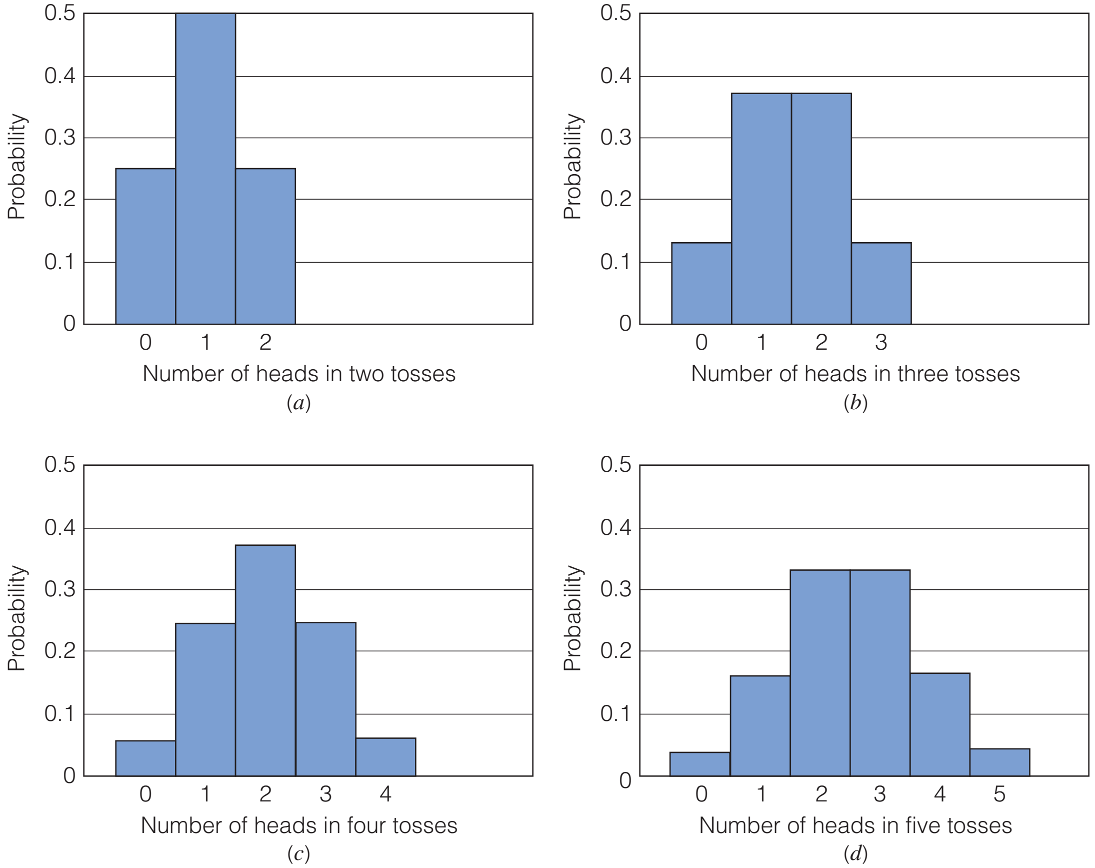


FIGURE 10.1 The binomial distribution for (*a*) $N = 2$, (*b*) $N = 3$, (*c*) $N = 4$, and (*d*) $N = 5$ when Pr (heads) = .5.

three combinations HHT, HTH, and THH, with a total probability of $1/8 + 1/8 + 1/8 = 3/8$. The resultant probabilities are shown in Figure 10.1*b*.

dichotomous observations observations that can be classified into only two categories

When tossing a coin, we can get only a head or a tail on any toss, and the result of one toss does not affect the next. When observations can be classified into just two categories like this, they are called **dichotomous observations**. Other examples include the results of a T-maze experiment in which rats must turn left or right and the results of a survey of voters in which respondents must state their preference for Jones or Smith. For a specified number of independent generating events in which there are only two possible events, the distribution that shows each possible outcome and the probability of each is called a **binomial distribution**.

binomial distribution when there are only two possible events, the distribution that shows each possible outcome and the probability of each

A binomial distribution may be interpreted as a **probability distribution** or, alternatively, as a **theoretical relative-frequency distribution** (see Section 2.5). It shows *the relative frequency with which certain outcomes occur in the long run*, that is, over an infinite number of trials. It does not mean that you will get exactly those outcomes in the short run. For example, look at the binomial distribution for tossing a fair coin three times in succession (Figure 10.1*b*). There are eight possible outcomes. The probability is .125 for TTT; .375 for TTH, THT, or HTT; .375 for HHT, HTH, or THH; and .125 for HHH. This does not mean that if we toss a coin three times, record the outcomes, and repeat the triple tossing seven more times, we would get exactly the distribution shown in Figure 10.1*b* (although that could happen). But if we were to keep recording the outcomes of the three tosses over a long series of triple tosses, and if we were to construct a relative-frequency distribution of our results, it would

probability distribution (a **theoretical relative-frequency distribution**) shows the relative frequency with which certain outcomes occur over an *infinite* number of trials

$(P + Q)^N$ the **binomial expansion;** formula to derive the binomial distribution

look very much like Figure 10.1*b*. For an infinite number of a series of three tosses, our relative-frequency distribution would look exactly like Figure 10.1*b* (assuming a fair coin fairly tossed).

Suppose we want to construct the binomial distribution for five tosses in a row. As the number of trials (in this case, successive coin tosses) increases, the number of possible outcomes increases dramatically (in fact, there are $2^5 = 32$ different possible outcomes for five successive coin tosses). Trying to figure out all the possible outcomes when the number of trials is five or greater is difficult without help. Fortunately, help comes in the form of a mathematical equation called the binomial expansion. For independent events, the **binomial expansion** is

$$(P + Q)^N \tag{10.1}$$

where: P = the probability of one of the events
Q = the probability of the other event $(= 1 - P)$
N = the number of trials in the series

The letters P and Q are standard symbols used to represent the probabilities of the two possible events. In our coin tossing example, we could just as easily use H and T, but the binomial describes a great number of distributions, not just coin tossing, so it is best to use the same symbols in all cases. Let us see how the binomial expansion can be used to derive the binomial distribution when we toss a coin two times. For this example, let P = probability of heads, Q = probability of tails, and $N = 2$. (Using the equally-likely model of probability, we have $P = Q = .5$.)

$$\begin{aligned}(P + Q)^2 &= P^2 + 2PQ + Q^2\\ &= (.5)^2 + 2(.5)(.5) + (.5)^2\\ &= .25 + .50 + .25\end{aligned}$$

Each term shows us a possible outcome and the exponent tells us how many heads or tails there are. Thus, P^2 represents two heads, Q^2 represents two tails, and $2PQ$ represents two possible outcomes of one head and one tail. The final probabilities of each are the same as those shown in Figure 10.1*a*.

For the example of three coin tosses, the binomial expansion gives us:

$$\begin{aligned}(P + Q)^3 &= P^3 + 3P^2Q + 3PQ^2 + Q^3\\ &= (.5)^3 + 3(.5)^2(.5) + 3(.5)(.5)^2 + (.5)^3\\ &= .125 + .375 + .375 + .125\end{aligned}$$

In this case, P^3 represents three heads in succession, $3P^2Q$ represents three possible outcomes of two heads and one tail, $3PQ^2$ represents three possible outcomes of one head and two tails, and Q^3 represents three tails in succession.

As you can see, except for a small number of trials, deriving different possible outcomes and the probabilities of each with the binomial expansion is also cumbersome.[1] Fortunately, there are tables based on the binomial expansion. Look at Table B in Appendix D. It gives the

---

[1] The equation to expand $(P + Q)^N$ is

$$(P + Q)^N = P^N + \frac{N}{1}P^{N-1}Q + \frac{N(N-1)}{1(2)}P^{N-2}Q^2 + \frac{N(N-1)(N-2)}{1(2)(3)}P^{N-3}Q^3 + \cdots + Q^N.$$

binomial distribution for values of N up to 15. For example, when $N = 3$ and $P = .5$, the probabilities are the same as those we derived earlier. To determine the probability of getting four heads in five tosses, we need only to look in the table to see that it is .1562.

10.5 Applying the Binomial

Example 1 Suppose that a psychologist is testing a rat using food reinforcement in a maze with six left–right choice points. At each choice point, one of the tunnels has been blocked (decided by tossing a coin). The psychologist first allows the rat to explore the maze and then tests the animal. If the rat makes five correct choices, what is the probability that it did so by chance (i.e., no learning)?

In the absence of learning, the probability of the rat choosing the correct side (P) at any choice point would equal 1/2, the same as the probability of making an incorrect choice (Q). Here $P = Q = 1/2$ and $N = 6$. To answer this question, let us expand the binomial:

$$\begin{aligned}(P + Q)^6 &= P^6 + 6P^5Q + 15P^4Q^2 + 20P^3Q^3 + 15P^2Q^4 + 6PQ^5 + Q^6\\ &= (.5)^6 + 6(.5)^5(.5) + 15(.5)^4(.5)^2 + 20(.5)^3(.5)^3 + 15(.5)^2(.5)^4\\ &\quad + 6(.5)(.5)^5 + (.5)^6\\ &= .0156 + .0938 + .2344 + .3125 + .2344 + .0938 + .0156\end{aligned}$$

We could, of course, have used Table B in Appendix D to obtain this same result. We see that the probability of exactly five correct choices is $6P^5Q = .0938$. The probability of six correct choices is $P^6 = .0156$. Thus, if the rat had not yet learned anything, the probability of five *or* six correct choices is $.0938 + .0156 = .1094$, or roughly 1 in 11.

Example 2 An instructor gives a 10-item multiple-choice quiz. Each question has five alternatives. If the student knows nothing about the subject and guesses on each question, what is the probability of him or her passing if the lowest passing score is six correct?

Let us assume that for each question, the choices are equally likely. If we let $P =$ the probability of a correct guess on any question, then $P = .20$. Looking in Table B in Appendix D for $N = 10$ and $P = .20$, we see that the probability of the student answering six or more questions correctly is .0055 (6 correct) + .0008 (7 correct) + .0001 (8 correct) + .0000 (9 correct) + .0000 (10 correct) = .0064.

Example 3 Suppose that a friend claims to have a homemade remedy for the common cold—a concoction made from chicken broth, vitamins, and some "secret" ingredients. We are skeptical and wish to put the remedy to a test. To do so, we find 16 other friends who are presently suffering with colds and match them in terms of severity of symptoms, forming eight pairs. In some pairs, both individuals have bad colds, and in other pairs both have moderate symptoms, but the two persons of any given pair have nearly equal symptoms. One member of each pair receives the remedy for 24 hours (in addition to his or her usual diet) and the other receives a placebo (water with artificial chicken flavoring). We determine which member of each pair receives the remedy by tossing a coin. We then find two friends in medical school to judge the subsequent severity of symptoms (neither of them knows which concoction was administered).

There are only two possible outcomes for each pair if ties are not allowed: the person taking the concoction is more improved or the other person is better. If the concoction does not work, for each pair the probability of the person taking the concoction being better is 1/2 (the same as obtaining heads on a toss of a coin). Let us suppose that at the end of 24 hours, seven of the eight persons who consumed the concoction are judged to be better than their partners. If the concoction has no healing powers, what is the probability of this outcome? Looking in Table B in Appendix D, we see that the probability is .0312. What is the probability that *at least* seven of the eight persons taking the concoction will be better than their partners? From the same table, we see that the probability of all eight being better is .0039; thus the probability of seven *or* more being better is .0312 + .0039 = .0351.

Theoretical probability distributions are a crucial part of inferential statistics. It is important that you understand the concept as fully as possible. What does it mean to say that the probability of at least seven of the eight persons who took the homemade remedy being better is .0351? *If the homemade concoction really has no healing powers*, and we were to repeat the test by observing the eight pairs again and again for an infinite number of times, on only 3.51% of the eight-pair tests would seven *or* eight persons taking the concoction be found to be better than the partner not taking the remedy. Thus, such an outcome would occur by chance less than 1 in 20 times if the concoction does not work. If seven of eight persons taking the concoction actually did get better, we would begin to suspect that the concoction really does have healing powers. But we are getting ahead of ourselves here. We will return to this type of problem in Chapter 12.

10.6 Probability and Odds

In previous sections, you computed probabilities. Recall that probabilities range from 0 to 1. Odds are nonnegative. That is, they can range from 0 to ∞. To compute the odds (Ω is the symbol for the odds), the equation is

$$\Omega = \frac{P}{1 - P} \tag{10.2}$$

For example, because the probability of a particular horse winning a race is .25, the odds of winning are $\Omega_{\text{Win}} = .25/(1 - .25) = 1/3$. This means that for every one race that this horse wins, it loses three. You can take the reciprocal of the odds of the horse winning to obtain the odds of the horse losing. Thus, Ω_{Lose} is 3/1. This means that for every three races that this horse loses, it wins one. As you can see in Formula 10.2, probabilities and odds are closely related.

In the social and behavioral sciences, odds are frequently used. In particular, the ratio of two odds (called an odds ratio) plays a central role in statistical procedures that are used to investigate the odds of an outcome such as being diagnosed with a sexually transmitted infection (or not diagnosed), being hired for a new job (or not hired), or whether a new drug is effective in treating a disorder (or not effective) (see Hosmer & Lemeshow, 2000). Notice that these outcomes are dichotomous observations. In Section 20.11, we revisit the odds ratio.

10.7 Are Amazing Coincidences Really That Amazing?

What are the chances that:

> Someone will win a state lottery twice within 4 months?
>
> You will meet someone at a party with the same birthday as you?

Surely these are highly unlikely chance events—one-in-a-trillion coincidences. Actually, they are not. In fact, they are very probable.

Let us consider again the case of Evelyn Adams, the woman we met at the beginning of the chapter who won the New Jersey Lottery twice in a 4-month time span in 1986. The probability of Ms. Adams doing this was 1 in 17 trillion. This is easily obtained from the multiplication theorem. If there were 4,123,000 lottery tickets sold for each lottery, and Ms. Adams had purchased 1 ticket for each, the probability of her winning both was (1/4,123,000)(1/4,123,000), the same as for any other *specific* person who purchased 1 ticket in each lottery.

But the probability of *someone, somewhere* winning two lotteries in 4 months is a different matter altogether. Professors Diaconis and Mosteller (1989) calculated the chance of this happening to be only 1 in 30. It is an example of what they call the "law of very large numbers": "With a large enough sample, any outrageous thing is apt to happen" (*New York Times*, Feb. 27, 1990, p. C1). Professor Diaconis gives the example of the blade-of-grass paradox: "Suppose I'm standing in a large field and I put my finger on a blade of grass. The chance that I would choose that particular blade may be one in a million. But it is certain that I will choose a blade." With millions of people buying lottery tickets in the United States, it's a pretty sure bet that someone, somewhere will win a lottery twice in his or her lifetime. For example, Joseph Crowley won a $3 million jackpot in the Ohio Lottery and then moved to Florida, where he won the $20 million Florida Lotto jackpot. For one-in-a-million coincidences, we would expect that in the United States, which has a population of approximately 350 million, there will be 350 such coincidences every day—another example of the law of very large numbers.

What about the probability that any two people at a party will have the same birthday? This really isn't uncommon at all. Professor Diaconis calculates that if you are in a room with 22 other persons, the chances are about even that 2 of you will have a birthday on the same day: $1 - \Pr$ (nobody has the same birthday) $= 1 - (364/365 \times 363/365 \times \ldots)$, and that value approaches .5 as $n = 22$.

Are there implications for this application of probability when we conduct a scientific study? Suppose that a group of researchers has conducted clinical trials of several new drugs with thousands of patients. They discover that in a certain subgroup (e.g., elderly, overweight men who drink heavily), twice as many who take drug A die compared to those taking drug B. One might immediately conclude that drug A is dangerous for certain people. However, with thousands of people taking dozens of drugs, the law of very large numbers comes into play—the probability is high that just by chance in *some* subgroup, twice the normal number of deaths will occur. The job of statistics is to determine these probabilities in advance so that we can tell whether our own results are to be expected or are indeed truly amazing.

10.8 Summary

The probability of an event is defined as the relative frequency with which that event would be observed over an infinite number of trials, when each trial is conducted in the same manner. It is thus a proportion (with a value between 0 and 1) characterizing an infinitely long series of occurrences.

We will never be able to determine a probability through direct observation of an infinite number of trials. However, if we can assume that all possible outcomes of a generating event have an equal likelihood of occurrence, we can calculate the probability of an event as the number of possible outcomes characterized by that event divided by the total number of all possible outcomes.

When outcomes are mutually exclusive, the probability of occurrence of any one of several particular outcomes is the sum of their individual probabilities. We call this the addition theorem of probability. When generating events are independent, the probability of successive or joint outcomes is the product of their separate probabilities. We call this the multiplication theorem.

When observations are dichotomous (classified into just two categories), we can construct a distribution that shows each possible outcome and its probability for a specified number of trials. This binomial distribution is a probability distribution that shows the relative frequency with which particular outcomes would occur over an infinite number of trials. Possible outcomes can be calculated from the binomial expansion, $(P + Q)^N$, but for the sake of convenience, we provide a table for up to 15 trials.

Mathematical Notes

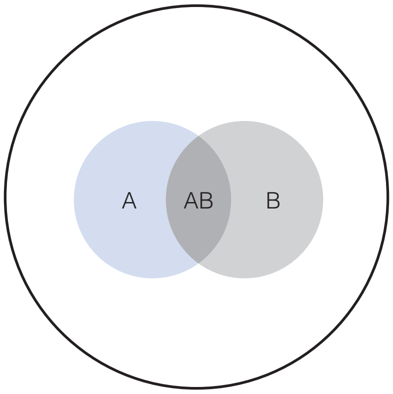
Note 10.1 Probability of Occurrence of A or B When A and B Are Not Mutually Exclusive (*Ref.*: Section 10.3)

$$\text{Pr (A or B)} = \text{Pr (A)} + \text{Pr (B)} - \text{Pr (AB)}$$

where: Pr (A) = the probability of A

Pr (B) = the probability of B

Pr (AB) = the probability of A *and* B occurring together



The outcome (A *or* B) is satisfied if A occurs without B, if B occurs without A, or, when A and B are not mutually exclusive, if both A and B occur. Note the accompanying illustration, in which A and B are *not* mutually exclusive. If the number of outcomes characterized by A is added to the number characterized by B, their sum includes *twice* the number characterized by the intersection of A and B. To determine the number of outcomes characterized by A or B, this total must therefore be reduced by the number characterized by A *and* B, so that the latter quantity will be counted only once. Similarly, the probability of A *or* B ought to include the probability of AB only once. When A and B *are* mutually exclusive, Pr (AB) = 0, and the formula for Pr (A or B) reduces to the addition theorem given in Section 10.3:

$$\text{Pr (A or B)} = \text{Pr (A)} + \text{Pr (B)}$$

Note 10.2 Conditional Probability (*Ref.*: Section 10.3)

conditional probability the probability of occurrence of A given that B has occurred

Pr (A|B) is the symbolic representation for the probability of occurrence of A *given that* B has occurred. (The vertical bar is read "given that.") Probabilities of this type are called **conditional probabilities**. Consider the selection of a card from a 52-card deck from which we have already drawn 1 card. According to the above definition, Pr (second card is an ace|first card is an ace) = 3/51, whereas Pr (second card is an ace|first card is not an ace) = 4/51.

The notion of conditional probability is important in defining the independence of events (see Note 10.3) and in considering sampling without replacement.

Note 10.3 Independence in a Probability Sense (*Ref.*: Section 10.3)

Read Note 10.2 first.

Section 10.3 stated that **independence** of events means that the outcome of one event must in no way alter the probability of occurrence of the other. Symbolically, we may say that A and B are independent if $\Pr(A) = \Pr(A|B)$. For example, in tossing coins, Pr (second coin is heads) = Pr (second coin is heads | first coin is heads), and the two events are independent. However, in a university, Pr (admission of an applicant) ≠ Pr (admission of an applicant | applicant's high school record is straight "A"); these two events are not independent.

Key Terms, Concepts, and Symbols

- event (p. 156)
- generating event (p. 156)
- relative-frequency definition of probability (p. 156)
- probability of an event (p. 156)
- Pr (p. 156)
- empirical probability (p. 157)
- equally-likely model of probability (p. 157)
- addition theorem of probability (p. 158)
- mutually exclusive (p. 158)
- multiplication theorem of probability (p. 159)
- independent events (p. 159)
- dichotomous observations (p. 160)
- binomial distribution (p. 160)
- probability distribution (p. 161)
- theoretical relative-frequency distribution (p. 161)
- binomial expansion (p. 161)
- conditional probability (p. 165)
- independence in a probability sense (p. 166)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. An instructor gives a 100-item multiple-choice examination. Each question has four alternatives. If the student knows nothing about the subject and selects answers by chance, (a) what is the probability for a given question that it will be answered correctly? (b) What is the expected number of correct answers for a given student (assuming he or she attempts all questions)?

2. An honest die is cast. What is the probability of rolling (a) a 5? (b) a 1 or a 6? (c) a 3 or higher? (d) on two rolls, a 2 followed by a 5? (e) on two rolls, a 2 followed by another 2?

3. Two honest dice are thrown. What is the probability of rolling (a) snake eyes—both 1s? (b) a 6 and a 6? (c) a 3 and a 4?

4. You toss a fair coin and roll an honest die. What is the probability of getting a head in combination with a 4?

5. From a deck of 52 cards, 2 cards are drawn at random. You replace the first card before drawing again. What is the probability that (a) both cards will be the jack of diamonds? (b) both cards will be jacks? (c) both cards will be diamonds? (d) the first card will be an ace and the second a king? (e) one card will be an ace and the other a king?

6. From a deck of 52 cards, 2 cards are drawn at random, but you do not replace the first card before drawing the second. What is the probability that (a) both cards will be the jack of diamonds? (b) both cards will be jacks? (c) both cards will be diamonds? (d) the first card will be an ace and the second a king?

7. Do you think that weather predictions ("the chance of rain tomorrow is 3 in 10") are better considered as theoretical probabilities or empirical probabilities? Explain.

8. Two honest dice are tossed. (a) List the 36 ways the dice may fall (remember to distinguish between the outcome of die #1 and the outcome of die #2). (b) Consider the sum of the points on the two dice; they may be from 2 to 12. With what relative frequency does each possible sum occur? (c) What is the probability of obtaining a 1 with the first die and a 3 with the second die? (d) What is the probability of obtaining a sum of points equal to 4? (e) What is the probability of obtaining a sum of points equal to 7 on one toss? (f) What is the probability of obtaining a sum of points greater than 7? Greater than 4? (g) What is the probability of obtaining a sum of points equal to 7 on two successive throws? Of 2 points on two successive throws?

9. You and five of your friends toss six fair coins at the same time. Construct the binomial distribution for all the possible results. What is the probability (a) that all six coins will come up tails? (b) that five of the six coins will come up tails? (c) that four of the six coins will come up tails?

10. You toss a fair coin 15 times. What is the probability of getting (a) 2 heads and 13 tails? (b) 5 heads and 10 tails? (c) 12 heads and 3 tails? (d) 10 or more heads?

11. A con artist uses a coin with a shaved edge to separate fools from their money. He bets that he can get at least 7 heads in 10 tosses. If altering the coin has made the Pr (heads) = .75, what is the probability of getting (a) 7 heads in 10 tosses? (b) at least 7 heads in 10 tosses?

12. A biology instructor gives a quiz consisting of 15 true–false items. She announces that the grades will be assigned as follows: A (14 or 15 correct); B (12 or 13 correct); C (10 or 11 correct); D (9 correct); F (8 or fewer). If a student has not studied for the quiz and just randomly guesses on each question, what is the probability that he or she will get a grade of: (a) A? (b) B? (c) C? (d) D or F? (e) C or better?

13. Fifteen shoppers are asked to participate in a taste test of four soft drinks at the local supermarket. Nine of them choose Brand C. If there were really no difference in taste among the four soft drinks, what would the probability be of 9 or more of 15 people choosing Brand C?

14. The Pop-up Toaster Company advertises that their toasters have a probability of .95 of working for 1 year without need of repairs. If you bought eight toasters as wedding gifts, what is the probability that all eight will work for at least 1 year?

15. Your friend reads in the paper about Professor Diaconis's conclusion that the chance of someone winning two state lotteries in 4 months is about 1 in 30. He wants to run out and buy 30 tickets in your state's lottery and another 30 in the neighboring state's lottery. If there are 3 million tickets sold in each lottery, what are his chances of winning both? How do you explain the discrepancy between what you have told him and the conclusion reached by Professors Diaconis and Mosteller?

CHAPTER 11

Random Sampling and Sampling Distributions

When you have finished studying this chapter, you should be able to:

- Define and explain random sampling and differentiate between sampling with replacement and sampling without replacement;
- Use a table of random numbers;
- Understand that the random sampling distribution of $\overline{X}$ is a (theoretical) probability distribution that shows the variability in the values of $\overline{X}$ for samples of the same size randomly drawn from a population;
- Determine the mean and standard deviation of the sampling distribution of $\overline{X}$;
- Explain the central limit theorem; and
- Use the sampling distribution of $\overline{X}$ to answer questions about the probability of obtaining $\overline{X}$s within a specified range of values.

Dr. Brown, the research director of a large school district, wants to know whether the sixth-grade students in her district are performing at the same level on a mathematics achievement test as students do nationally. In her district, there are 2,500 sixth-grade students, too many to study conveniently. She therefore selects 100 students from among the 2,500, finds their mean mathematics test score, and compares it with the mean of the nationwide group. If the mean of her group differs from the national mean, what should she conclude?

In the chapters to come, you will learn how to draw conclusions in problems like the one just presented. It is a problem in statistical inference. Recall from Chapter 1 that a basic aim of **statistical inference** is to form a conclusion about a population **parameter** from a sample (**statistic**) taken from that population.

A **population** is the complete set of observations about which an investigator wishes to draw conclusions. A **sample** is a part of that population. Notice that *a population is defined in terms of observations rather than people*. This way of looking at things works better in statistics. Suppose, for example, that we want to know both the SAT scores and the high school grade point averages for the entering freshmen in our college. We are talking about two populations, despite the fact that they are attached to the same people. For convenience, and to avoid awkward language, we often say "a sample of college freshmen" rather than "a sample of SAT scores obtained from college freshmen." But statistically speaking, both sample and population consist of measurements (observations) rather than the subjects from which they were derived.

Notice also that *a population is defined by the interest of the investigator*. If pollsters are trying to determine the voter preferences of the nation prior to an election, their target population consists of the preferences of all those who will go to the polls and vote (over 100 million). However, sometimes observations obtained from the group at hand are just what we want, and therefore they would constitute the population. For example, if an instructor wants to know how the present class performed on the first midterm, the students' scores are the population and not a sample. In our present example, the 2,500 sixth-grade test scores from Dr. Brown's district are the population.

In inference, we often use a procedure called hypothesis testing. The investigator first poses a research question (see Section 1.3). However, the procedure for answering a statistical question is easier if it is translated into a **hypothesis**, a statement about a population parameter to be subjected to test and, on the outcome of the test, to be retained or rejected. In our example, Dr. Brown will hypothesize that the mean of the mathematics achievement test in her school district (her *population*) is equal to the mean of the national group. *To evaluate her hypothesis, Dr. Brown will ask what type of sample results are likely to occur if the hypothesis were correct. If her sample outcome is unlikely, she will reject the hypothesis.*

In our example problem, the data at hand are not those about which Dr. Brown really wishes to draw her conclusion. She wants to know how all 2,500 sixth-grade students in her school district score on the mathematics achievement test, not just the 100 she selected. In short, Dr. Brown wants to draw a conclusion about the population represented by the sample she studied and not simply about the particular sample itself.

In inference, the fundamental factor that must be taken into account is that the value of the characteristic we are studying will vary from sample to sample. Suppose Dr. Brown replicates her study several times. On each repetition she will draw a different selection of 100 subjects from the 2,500 sixth-grade students in her district. Even if sixth-grade students in her school district score as well on the mathematics achievement test as students do nationally, the mean scores from different samples will differ from the national mean (and from one another) due to chance factors involved in sampling.

The key to any problem in statistical inference is to discover what sample values will occur by chance in repeated sampling and with what probability. We draw a sample and compute the mean. If we draw a second sample of the same size it is likely that the mean would have a different value. A third is likely to give still another value. What kind of distribution will be formed by the means when the sampling is repeated again and again? A distribution of this type is known as a **sampling distribution**. We must be able to describe it completely if we are to say what would happen when samples are drawn.

statistical inference
to draw a conclusion about a population parameter from a sample taken from that population

parameter
a descriptive index of a population

statistic
a descriptive index of a sample

population
the complete set of observations about which an investigator wishes to draw conclusions

sample
a subset of a population

hypothesis
a statement about a population parameter to be subjected to test and, on the outcome of the test, to be retained or rejected

sampling distribution
a theoretical relative frequency distribution of the values of a statistic that would be obtained by chance from an infinite number of samples of a particular size drawn from a given population

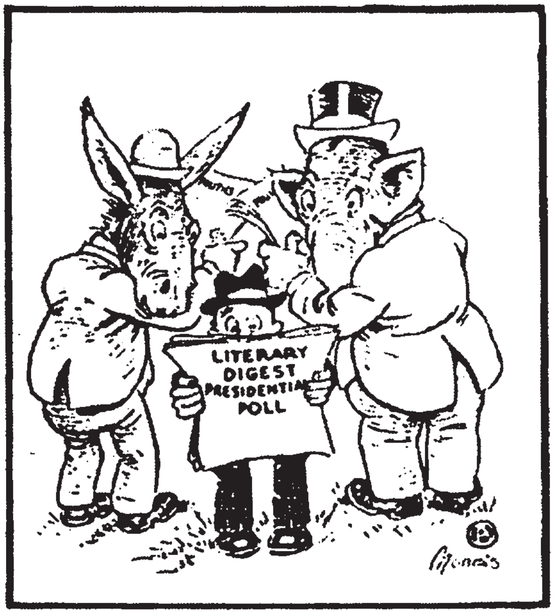
To learn what sample values will occur and with what probability, there must be known rules that connect the "behavior" of the samples to the population from which they are drawn. We can know these rules only if a systematic method of drawing samples is adopted and used *consistently*. As it turns out, there is just one basic method of sampling that permits the rules to be known. This is to draw **probability samples**—samples for which the probability of inclusion in the sample of each element of the population is known. One kind of probability sample is the random sample; it is the kind we consider in this book.

probability samples samples for which the probability of inclusion in the sample of each element in the population is known

11.1 Random Sampling

There is nothing in the definition of a sample that specifies how it is to be selected from a population. However, the method of selection matters very much in statistics. For example, in the 1936 American presidential election, the incumbent, Franklin D. Roosevelt, a Democrat, was being challenged by Alf Landon, a Republican. One of the leading magazines of the day, the *Literary Digest*, surveyed voter preferences by mailing questionnaires to 10 million people whose names were gathered from lists of automobile and telephone owners. Over 2 million people responded, and the results indicated that Landon would beat Roosevelt by a landslide. In fact, Roosevelt beat Landon by one of the largest margins ever. This was one of the largest surveys ever taken. How could it have been so wrong? The United States was in the middle of the Great Depression in 1936 and only a minority of people were financially secure enough to own a car and a telephone. They tended to vote Republican. Most other Americans were worried about buying enough food to feed their families, and they tended to vote Democratic. In other words, the *Literary Digest* had surveyed an unrepresentative subgroup of the population (Kaplan & Saccuzzo, 1989).

If it is not the size (alone) of a sample that determines whether or not it is a good sample, what does? An essential assumption in inferential statistics is that samples are drawn randomly



By Morris From George Matthew Adams Service

Oh! Oh!

Political cartoon from the 1930s.

random sample
a sample obtained in a way that ensures that all samples of the same size have an equal chance of being selected from the population

from a particular population. A **random sample** of a given population is a sample so drawn that each possible sample of that size has an equal probability of being selected from the population. The method of selection, not the particular sample outcome, defines a random sample. If we were given a population and a sample from that population, it would be impossible to say whether the sample was random without knowing the method by which it was selected. Essentially, a random sample is one where chance does the selecting, as in a properly conducted lottery.

Most people believe that a random sample is always representative of the population from which it is drawn, but that is not necessarily so (Tversky & Kahneman, 1971). Although characteristics of random samples *tend* to resemble those of the population, those of a particular random sample may not. If we deal 13 cards from a well-shuffled 52-card deck, each of the four suits (clubs, hearts, diamonds, spades) will usually be represented. Somewhat infrequently, a hand of 13 cards will contain cards from only three suits, and even less frequently, from only two suits. There are two properties of random samples that are very important for you to understand. First, if several random samples are drawn from the same population, the samples will almost always differ, and therefore their characteristics (statistics) will vary from sample to sample. Second, we would expect that the larger the *random* samples from a given population, the less the variation in characteristics from sample to sample. We will return to these two properties later in this chapter.

sampling with replacement
a sampling procedure in which an element may appear more than once in a sample

sampling without replacement
a sampling procedure in which no element may appear more than once in a sample

Although there is only one way to define a random sample, there are two sampling plans that yield a random sample: **sampling with replacement** (in which an element may appear more than once in a sample) and **sampling without replacement** (in which no element may appear more than once). We discuss and compare these sampling plans in Section 11.6. However, until then, assume that we are using the *with-replacement* plan.

To obtain a random sample, *every element in the population must have an equal chance of being included in the sample*. (Note that this satisfies the basic assumption of the equally-likely model of probability—see Section 10.2.) However, in the real world of data, it is often difficult or costly (and sometimes impossible!) to draw a simple random sample from the population that we would truly like to study. Sometimes we are tempted to use a shortcut to reduce the amount of work. A common method is to select what might be called a *casual sample*. The essence of this method is that the investigator attempts to act as a human randomizer. He or she tries to choose "at random" from the available subjects, but the method will more likely fail than succeed. For example, what about assigning rats from the new shipment to two conditions of an experiment? Our research assistant reaches into the shipping containers to select those animals to receive Treatment A. Do you think animated and lethargic rats will have an equal chance of being chosen? What about those that bite at fingers versus those that do not? Take it from Minium's Third Law of Statistics: "*Human beings are rotten randomizers*."

As an alternative, we could write the number of each rat on a ticket, place the tickets in a can, shake the can thoroughly, and make a blindfolded selection of subjects for the two groups. An even better method is to use a table of random numbers to select the sample (or use a computer program that accomplishes the same thing).

11.2 Using a Table of Random Numbers

Because human judgment cannot be depended upon to select at random, how should we do it? If each element of the population can be identified and assigned a different number, a convenient and reliable procedure is to use a **table of random numbers**, such as Table C in Appendix D. Such tables are usually constructed by computer, according to a program that ensures each digit from 0 to 9 has an equal probability of occurrence.

A table of random numbers may be read as successive single digits, as successive two-digit numbers, as three-digit numbers, and so on. We may also read the numbers vertically (in either direction) or horizontally (in either direction). For example, in Table C, if we begin at the top left on the first page and *read down*, the random order of single-digit numbers would be 1, 9, 0, 7, 3, . . . If we wish to read two-digit numbers, they would be 11, 96, 07, 71, 32, . . . To use the table properly, we must adopt some scheme *in advance*, including how to enter the table and how to proceed when the end of a column or row has been reached.

As an example, let us suppose that 30 scores are to be selected at random from a population of 700. To begin, assign each element of the population an identification number. The order in which you number the elements does not matter because subsequent random sampling will eliminate any bias. Each element may be given a three-digit identification number from 001 to 700. For the sake of simplicity, suppose that our scheme requires that we start at the upper left of the first page of the table and read down. (We could start anywhere else.) The first random number is 113, and the first element to be included in the sample is therefore the element bearing the identification number 113. The next random number is 969; we skip it because it identifies no element of the population. The next number is 077, which identifies the second score to be included in the sample. We continue until 30 elements have been selected.

If our sampling is *with replacement* and the same element is selected again, we must include it again. If our sampling is *without replacement* and an identification number appears again, we skip it.

11.3 The Random Sampling Distribution of the Mean: An Introduction

Suppose there is a very large population of scores (X) with a mean, μ_X, equal to 100. We now draw from this population a random sample of 25 cases; its mean, $\overline{X}_1$, proves to be 97.4. We now replace the 25 cases in the population, "stir well," and select a second sample of 25 cases at random. Its mean, $\overline{X}_2$, is 101.8. Theoretically, we could repeat this sampling procedure indefinitely, and for each sample calculate and record the mean. Figure 11.1 illustrates the situation.

If we continued these sampling trials for an infinite number of trials, and if we then put all the sample means into a relative frequency distribution, what sample values will occur in repeated sampling, and with what probability (relative frequency)? The distribution of sample means

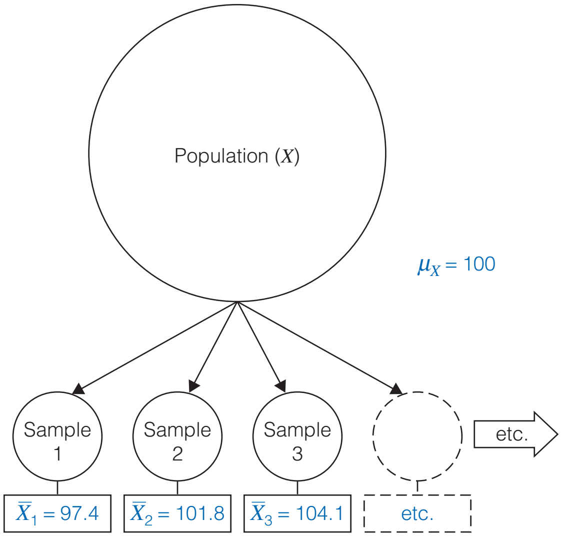


FIGURE 11.1 Means of successive random samples.

random sampling distribution of the mean ($\overline{X}$) a theoretical relative frequency distribution of all the values of $\overline{X}$ that would be obtained by chance from an infinite number of samples of a particular size drawn from a given population

that would be generated is known as a random sampling distribution of the mean. A **random sampling distribution of the mean** is the relative frequency distribution of $\overline{X}$ obtained from all possible random samples of a given size that could be drawn from a given population.

Notice that the sampling distribution of the mean, like the binomial distribution (Chapter 10), is a *probability distribution*. Whereas the binomial distribution shows all the possible outcomes for dichotomous observations (such as tossing a coin) and the probability of each, a sampling distribution of the mean shows all the possible values that the sample mean can take on (for samples of a given size) and the probability of each.

If the population is of finite size, the number of possible samples will also be finite. This makes it possible to generate an entire sampling distribution and to explore its properties. However, to do this we will have to keep things simple. Consider a population of four scores—2, 4, 6, and 8—from which we are to select samples of size 2. What are the possible samples? When sampling is done *with replacement* (i.e., each score is drawn, recorded, and then returned to the population before the next score is chosen), there are four ways in which the first score may be chosen and also four ways in which the second may be chosen. Consequently, there are $4 \times 4 = 16$ possible samples. These are shown in the first column of Table 11.1.

What is the probability of drawing the sample (2,2)? When each score is given an equal opportunity of selection, the probability that the first element will be 2 is 1/4, and when the score is replaced before selecting the second element, the probability that the second element will be 2 is also 1/4. Selection of the second element is in no way dependent on the outcome of the first element, so the two events are independent. The probability, therefore, of the sample (2,2) is the product of the two probabilities, $(1/4)(1/4) = 1/16$, according to the multiplication rule for independent events (presented in Section 10.3). By similar reasoning, the probability of any one of the other samples is also 1/16. If samples are selected in this manner, each of the 16 possible samples is equally likely to occur, and the basic condition of random sampling is satisfied.

Our fundamental interest is in the means of these samples rather than in the samples themselves. The mean of each sample is given in the third column of Table 11.1. Note that

TABLE 11.1 *Possible Samples and Sample Means for Samples of Size 2 (Sampling with Replacement)*

POPULATION: 2, 4, 6, 8

Sample	*Probability of occurrence*	*Mean*
2,2	1/16	2.0
2,4	1/16	3.0
2,6	1/16	4.0
2,8	1/16	5.0
4,2	1/16	3.0
4,4	1/16	4.0
4,6	1/16	5.0
4,8	1/16	6.0
6,2	1/16	4.0
6,4	1/16	5.0
6,6	1/16	6.0
6,8	1/16	7.0
8,2	1/16	5.0
8,4	1/16	6.0
8,6	1/16	7.0
8,8	1/16	8.0

TABLE 11.2 *Sampling Distribution of the Mean (Data from Table 11.1).*

SAMPLE MEANS	RELATIVE FREQUENCY
8.0	1/16
7.0	2/16
6.0	3/16
5.0	4/16
4.0	3/16
3.0	2/16
2.0	1/16

although there are 16 different samples, each equally likely, there are not 16 different means, each equally likely. For example, there is only one sample (2,2) that yields the mean of 2.0, but there are two samples for which the mean is 3.0: (2,4) and (4,2).[1] The probability of obtaining a sample with a mean of 2.0 is therefore 1/16, whereas the probability of obtaining a sample with a mean of 3.0 is 1/8, twice as great. This illustrates an essential point: *Random sampling results in equal probability of all possible samples, not in equal probability of all possible sample means.*

Suppose we put the 16 means in a relative frequency distribution, as Table 11.2 shows. This relative frequency distribution is, of course, the random sampling distribution of $\overline{X}$ for samples of size 2 drawn with replacement from the particular population of size 4. This distribution is shown graphically in Figure 11.2. For comparative purposes, two features have been added. First,

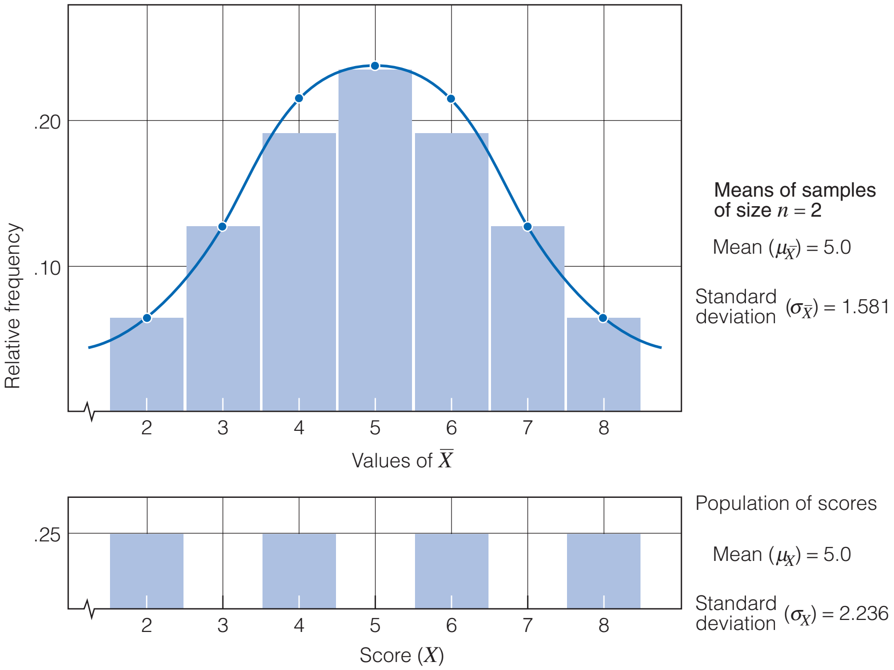


FIGURE 11.2 Distribution of means of the 16 possible samples of size 2 obtainable from a population of four scores. Data from Table 11.2. Population: 2, 4, 6, and 8.

---

[1] The sample (2,4) must be treated as distinct from the sample (4,2). Each is possible, and each yields a sample mean that must be accounted for, just as head-tail is different from tail-head.

the distribution of the population of scores appears in the same figure. Second, a normal curve has been fitted to the sampling distribution of the mean. What features appear in this sampling distribution?

11.4 Characteristics of the Random Sampling Distribution of the Mean

In most real-world examples, it is impossible to produce the actual relative frequency distribution of $\overline{X}$ from samples randomly drawn from a population because it calls for an *infinitely* long series of samples. However, mathematicians have been able to derive its characteristics and tell us what would happen *if* an infinite series of trials were conducted. It is the fact that the samples are selected *at random* that makes this possible.

expected value (of the sample mean) the mean of a random sampling distribution of $\overline{X}$

$\mu_{\overline{X}}$ symbol for the expected value of the sample mean; $\mu_{\overline{X}} = \mu_X$

Recall that a distribution is completely defined by its mean, standard deviation, and shape. Let us consider the mean first. The mean of any random sampling distribution of $\overline{X}$, called the **expected value** of the sample mean, is the same as the mean of the population of scores from which the samples were drawn. This is true regardless of n, σ, and the shape of the population. To put it symbolically:

MEAN OF THE SAMPLING DISTRIBUTION OF THE MEAN ($\overline{X}$)

$$\mu_{\overline{X}} = \mu_X \qquad (11.1)$$

Using Figure 11.2 as an example, $\mu_{\overline{X}} = \mu_X = 5.0$.

standard error of the mean the standard deviation of the random sampling distribution of $\overline{X}$

$\sigma_{\overline{X}}$ symbol for the standard error of the mean, $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$

The standard deviation of the random sampling distribution of the mean, called the **standard error of the mean**, depends on the standard deviation of the population, σ_X, and the sample size, n. Symbolically:

STANDARD ERROR OF THE MEAN

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \qquad (11.2)$$

For the sampling distribution in Figure 11.2, we see that $\sigma_{\overline{X}} = 2.236/\sqrt{2} = 1.581$, which is *less* than the standard deviation of the population of scores from which the samples were taken. The standard error of the mean is the standard deviation of the sampling distribution of $\overline{X}$ based on samples of a specified size. The term *standard error* is a vestige from the time when the normal curve was known as the "normal law of error" (see Section 6.1).

shape of the sampling distribution of $\overline{X}$ if the population of scores is normally distributed, the sampling distribution of $\overline{X}$ will also be normally distributed, regardless of sample size

Study Formula 11.2 carefully. It shows (a) that sample means vary less than scores do (when the sample size is at least 2), (b) that sample means vary less when scores in the population vary less, and (c) that sample means vary less when sample size is greater.

Note that Greek letters are used to specify the mean and standard deviation of the sampling distribution ($\mu_{\overline{X}}$, $\sigma_{\overline{X}}$). They remind us that the sampling distribution constitutes a population of values, not a sample of them. This distribution represents *all* the possible means of samples of a specified size and therefore constitutes the population of them—not just *some* of them.

With regard to the shape of the distribution, *if the population of scores is normally distributed, the sampling distribution of the mean will also be normally distributed, regardless of sample size.* If the population of scores is normally distributed, it does not matter whether we take an infinite

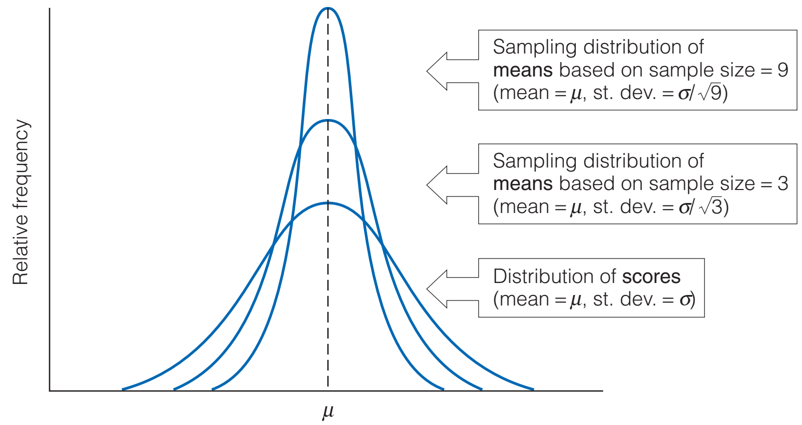


FIGURE 11.3 A normally distributed population of scores and random sampling distributions of means for $n = 3$ and $n = 9$.

number of samples of size 2 or an infinite number of samples of size 200. In both cases, the sampling distribution of $\overline{X}$ will also be normally distributed.

You can see these three properties of the sampling distribution of the mean in Figure 11.3, which shows two different sampling distributions (for samples of sizes 3 and 9) along with the normally distributed population of scores from which the samples were drawn. Note, in particular, how the sampling distribution changes with a change in sample size. This figure reminds us that there is not just one random sampling distribution of $\overline{X}$ corresponding to a given population, but a *family* of such distributions, one for each possible sample size.

central limit theorem see definition in text; regardless of the shape of the population of scores from which random samples are drawn, the sampling distribution of $\overline{X}$ more and more resembles a normal distribution as sample size increases

Now, all is well and good if the population is normally distributed, but what if it is not? Fortunately, a remarkable bit of statistical theory comes to the rescue: the **central limit theorem**. This theorem informs us that

> the random sampling distribution of the mean tends toward a normal distribution irrespective of the shape of the population of observations sampled; the approximation to the normal distribution improves as sample size increases.

With some populations, the distribution of scores is sufficiently similar to a normal distribution so that little assistance from the central limit theorem is needed. However, even when the distribution of the population of scores differs substantially from a normal distribution (see Point of Controversy in Chapter 6), the sampling distribution of $\overline{X}$ may be treated as though it were normally distributed when the sample size is reasonably large. How large is "reasonably large"? Depending on the problem, 25 to 30 cases is usually quite large enough.

Figure 11.4 gives some idea of the tendency of the sampling distribution of $\overline{X}$ to resemble the normal curve as sample size rises. Figure 11.4*a* shows two populations of scores, one rectangular, one skewed. Figure 11.4*b* shows sampling distributions of $\overline{X}$ for samples based on $n = 2$. Note that the shapes of these distributions differ from those of their parent populations of scores and that the difference is in the direction of a normal curve. Figure 11.4*c* shows the sampling distribution of $\overline{X}$ of samples based on $n = 25$. In both cases the degree of resemblance to the normal distribution is remarkable, though the resemblance is closer for means of samples drawn from the rectangular distribution. The truth of the central limit theorem is beautifully illustrated in Figure 11.2. Look at the distribution of population scores at the bottom of the figure. This certainly is no normal distribution; if it had to be given a name, we might call it the "tooth distribution of the Halloween pumpkin." But look at the random sampling distribution of the mean. It comes very close to a normal distribution. Compare the heights of the rectangles with the dots showing the elevation of the normal distribution at similar locations.

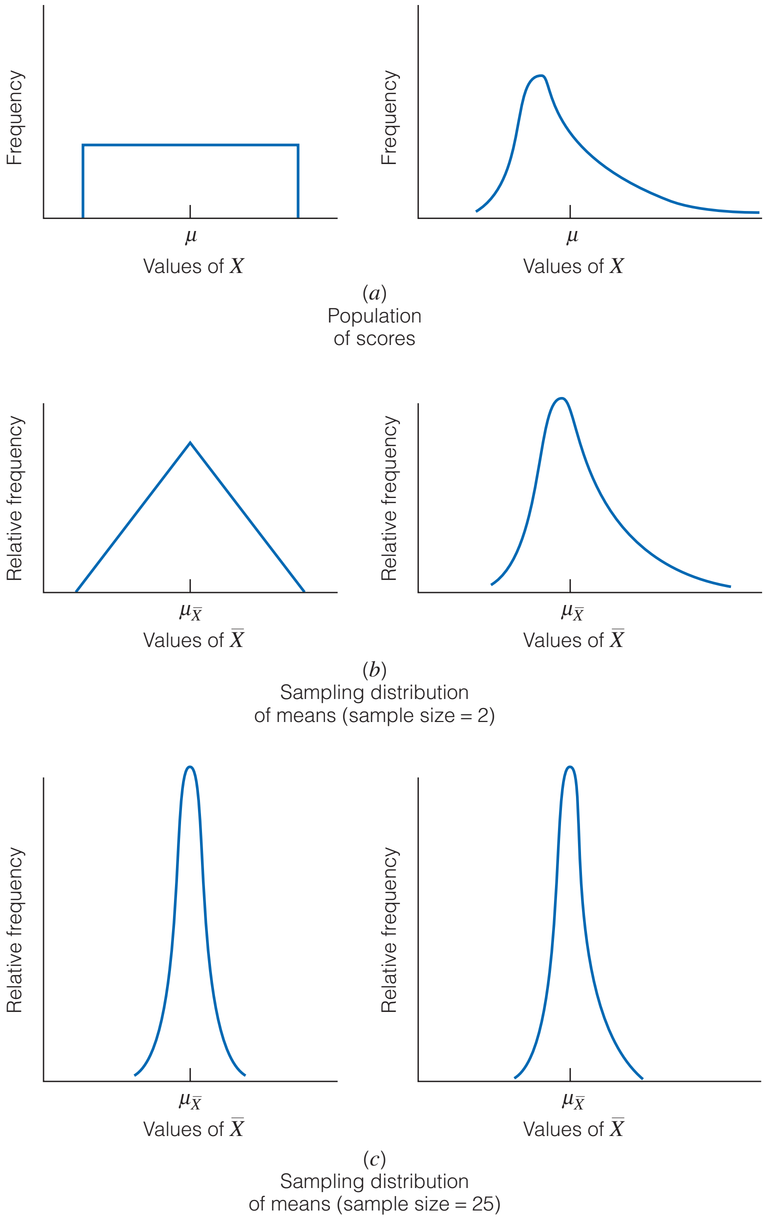


FIGURE 11.4 Effect of the central limit theorem.

11.5 Using the Sampling Distribution of $\overline{X}$ to Determine the Probability for Different Ranges of Values of $\overline{X}$

We may use the random sampling distribution of the mean to determine the probability with which sample means would fall between certain limits. In Chapter 6, you learned that many useful questions can be answered when we have a normal distribution of known mean and standard deviation. In this section, you will learn to use those procedures to answer four common types of questions about the sampling distribution of $\overline{X}$. Be sure to review the methods of Chapter 6 (Sections 6.3–6.5) before proceeding.

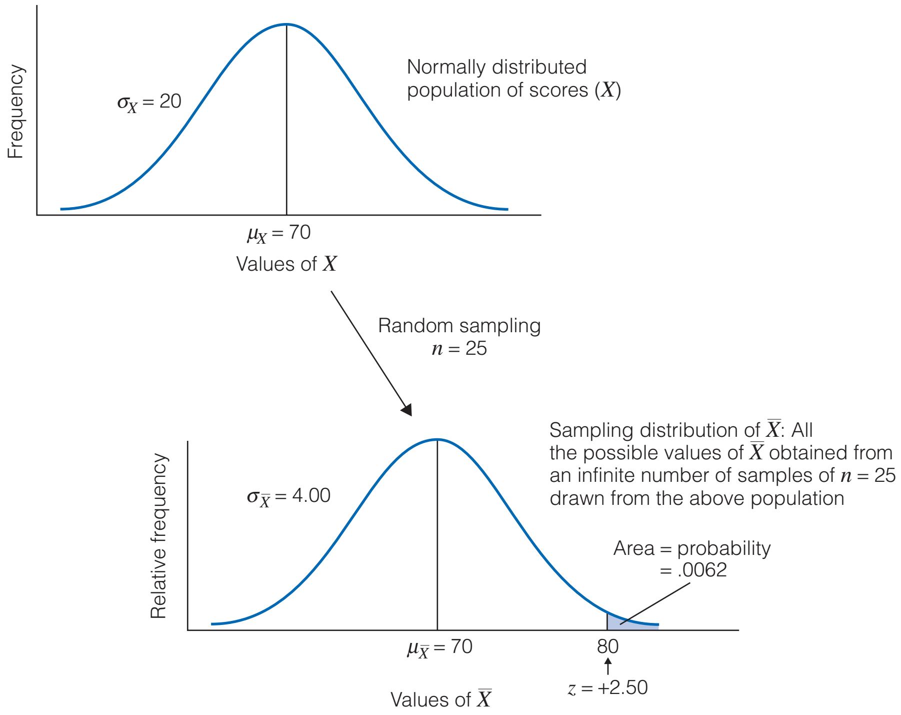

Frequency
$\sigma_X = 20$
Normally distributed population of scores (X)
$\mu_X = 70$
Values of X
Random sampling
$n = 25$
Relative frequency
$\sigma_{\overline{X}} = 4.00$
Sampling distribution of $\overline{X}$: All the possible values of $\overline{X}$ obtained from an infinite number of samples of $n = 25$ drawn from the above population
Area = probability = .0062
$\mu_{\overline{X}} = 70$
80
$z = +2.50$
Values of $\overline{X}$

FIGURE 11.5 Finding the proportion of sample means exceeding a given value.

Given

A normally distributed population, with $\mu = 70$ and $\sigma = 20$. Assume that your sample size is 25.

Problem 1

What is the probability of obtaining a random sample with a mean of 80 or higher?

Figure 11.5 illustrates the sampling distribution of $\overline{X}$ for an infinite number of samples of $n = 25$ drawn from a normally distributed population with $\mu_X = 70$ and $\sigma_X = 20$. Because the population of scores is normally distributed, the sampling distribution of $\overline{X}$ is also normally distributed. We also know that the sampling distribution has a mean (expected value) of 70 (remember, $\mu_{\overline{X}} = \mu_X$).

Solution

Step 1: Calculate the standard deviation of the sampling distribution, that is, the standard error of the mean (Formula 11.2).

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{20}{\sqrt{25}} = 4.00$$

z-score formula for a value of $\overline{X}$ (in a sampling distribution of $\overline{X}$)

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

Step 2: Translate your sample mean of 80 to a z score. The general formula for a z score is (*score* − *mean*)/(*standard deviation*). In a sampling distribution of $\overline{X}$, the sample mean is the score; the mean of the sampling distribution (which is equal to the mean of the population) is the mean; and the standard error of the mean is the standard deviation. Therefore,

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{80 - 70}{4} = +2.50$$

Step 3: Take the z score to the table of areas under the normal curve (Table A in Appendix D), and determine the proportionate area beyond z. It is .0062. Now, because proportion

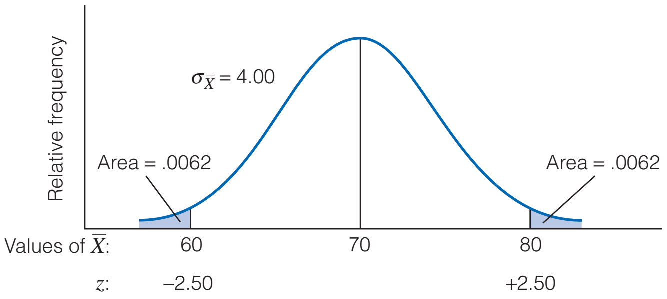


FIGURE 11.6 Finding the proportion of sample means that differ from the population mean by more than a given amount.

of area under the curve corresponds to proportion of scores in the distribution (see Section 6.3), you can see that in random sampling, means of 80 or higher would occur with a relative frequency of .0062 for samples of size 25. Therefore, the probability of obtaining such a mean or larger in random sampling from the given population is .0062.

Problem 2

What is the probability of obtaining a sample with a mean that differs from the population mean by 10 points or more?

Solution

In Problem 1, the mean of 80 fell 10 points above the expected value, the mean (70) of the sampling distribution. Because the normal curve is symmetrical, the same area lies below a mean of 60 (10 points below 70) as lies above a mean of 80 (10 points above 70). The z scores of both will have the same magnitude; we must double the corresponding area to find the required probability. In random sampling, assuming samples of size 25, the probability of obtaining a sample mean differing from the population mean by 10 points or more is therefore $(2)(.0062) = .0124$. The problem is illustrated in Figure 11.6.

Problem 3

What mean has such a value that the probability is .05 of obtaining one as high or higher in random sampling? This problem is illustrated in Figure 11.7.

Solution

Step 1: From Column 3 in Table A of Appendix D, find the z score above which .05 of the area under the normal curve falls; it is between $z = 1.64$ and $z = 1.65$, or about $+1.645$.

Step 2: The value of z informs us that the desired value is 1.645 standard deviations above the mean of the distribution. Therefore (remembering that our concern is with the sampling distribution of $\overline{X}$ and that its standard deviation is $\sigma_{\overline{X}}$, not σ_X):

$$\begin{aligned}\overline{X} &= \mu_{\overline{X}} + z\sigma_{\overline{X}} \\ &= 70 + (1.645)(4.00) \\ &= 70 + 6.58 \\ &= 76.58\end{aligned}$$

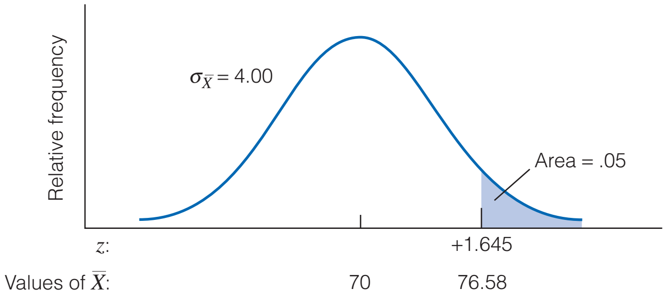


FIGURE 11.7 Finding the value above which a given proportion of sample means will fall.

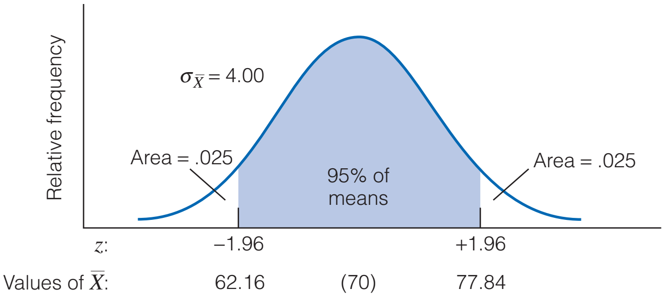


FIGURE 11.8 Finding the centrally located score limits between which a given proportion of sample means will fall.

For this example, 5% of the sample means will have a value of at least 76.58; the probability is .05 of obtaining such a mean.

Problem 4 Within what limits would the central 95% of the sample means fall?

Solution If .95 of the values are to fall in the center, the remaining .05 must be divided equally between the two tails of the distribution. Therefore, we must find the value of z beyond which .025 of the sample means are located. The method is the same as that for Problem 3, and the solution is outlined as follows:

Lower Limit	**Upper Limit**
$z_{\text{LL}} = -1.96$	$z_{\text{UL}} = +1.96$
$\overline{X}_{\text{LL}} = \mu_{\overline{X}} + z_{\text{LL}}\sigma_{\overline{X}}$	$\overline{X}_{\text{UL}} = \mu_{\overline{X}} + z_{\text{UL}}\sigma_{\overline{X}}$
$= 70 + (-1.96)(4.00)$	$= 70 + (+1.96)(4.00)$
$= 70 - 7.84$	$= 70 + 7.84$
$= 62.16$	$= 77.84$

For this example, 95% of the sample means fall between 62.16 and 77.84. In random sampling, the probability is .95 of obtaining a mean within these limits and .05 of obtaining one beyond these limits. The problem is illustrated in Figure 11.8.

Problems like the four preceding are more than just mathematical exercises. Starting with the next chapter, we will use procedures like these to answer real research questions.

11.6 Random Sampling Without Replacement

If there are 50 tickets in a lottery and we draw one ticket, set it aside, draw another, and continue until five tickets have been selected, we are *sampling without replacement*. Under this sampling plan no element may appear more than once in a sample. On the other hand, if we select a ticket, note its number, and return it before choosing the next ticket, we are *sampling with replacement*. Under this plan it is possible to draw a sample in which the same element appears more than once. Both of these plans satisfy the condition of random sampling, but *certain sample outcomes that are possible when sampling with replacement are not possible when sampling without replacement*.

Formula 11.2 assumes sampling with replacement. The standard error of the mean is smaller when sampling without replacement. However, *the difference is trivial if the sample is small relative to the population*.

Despite the fact that most sampling in behavioral science is done without replacement, we typically use Formula 11.2 for the standard error of the mean because although we do not know N, it is reasonable to assume that it is large.

11.7 Summary

A sampling distribution of a statistic is a probability distribution. It shows all the values of a statistic for an infinite number of samples of a given size drawn from a particular population. Some values are more probable than others. To construct a sampling distribution, a method of drawing samples must be specified and used consistently. Random sampling is the fundamental method that makes it possible to discover the properties of the sampling distribution. For the random sampling distribution of $\overline{X}$, the mean of the infinite number of $\overline{X}$s, $\mu_{\overline{X}}$, called the expected value, is equal to μ_X, the mean of the population of observations from which the samples were taken. Its standard deviation, $\sigma_{\overline{X}}$, called the standard error of the mean, is equal to $\sigma_X/\sqrt{n}$. Thus, as sample size increases, the standard error of the mean becomes smaller (i.e., the sample means cluster closer to the true population mean). If the population of observations is normally distributed, the sampling distribution of $\overline{X}$ also will be normally distributed. If not, the central limit theorem tells us that the normal distribution will often still be a close, and therefore usable, approximation if sample size is not too small. If we know these three characteristics, we can determine the probability of obtaining a sample mean (drawn from a specified population) with values as extreme or more extreme than the one we have drawn.

Key Terms, Concepts, and Symbols

statistical inference (p. 169)
parameter (p. 169)
statistic (p. 169)
population (p. 169)
sample (p. 169)
hypothesis (p. 169)
sampling distribution (p. 169)
probability samples (p. 170)
random sample (p. 171)
sampling with replacement (p. 171)
sampling without replacement (p. 171)
table of random numbers (p. 171)
random sampling distribution of the mean (p. 173)
three properties of a sampling distribution of $\overline{X}$ (p. 175)
expected value (p. 175)
$\mu_{\overline{X}}$ (p. 175)
standard error of the mean (p. 175)
$\sigma_{\overline{X}}$ (p. 175)
shape of the sampling distribution of $\overline{X}$ (p. 175)
central limit theorem (p. 176)
sampling distribution as probability distribution (p. 177)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. You are a "person-in-the-street" interviewer for a radio station. Suppose that your population consists of those who come close enough that you might ask them for an interview. Name four biases that you think might prevent you from obtaining a truly random sample of interviewees.

2. In recent years, shopping malls have become a favorite site for companies to test consumers' responses to new products through interviews. These interviews, called "mall intercepts," have largely replaced door-to-door interviewing because they are relatively quick and inexpensive. Research has shown that 95% of American households shop at malls at least once a year and that two-thirds shop at malls at least once every 2 weeks. (a) Can consumers interviewed in this manner be considered a random sample of the population as a whole? Explain. (b) Is the sample a random sample of people who go to malls? Explain. (c) Is it a random sample of people who go to that mall? Explain.

3. Consider the scores in Data 2A of Problem 8 in Chapter 2. Using the table of random numbers (Table C in Appendix D) and the without-replacement sampling plan, select a sample of 10 scores from this set of 80 scores. Describe the steps in your procedure, including how you used the table of random numbers.

4. Consider Table 11.2. In random sampling, what is the probability of obtaining a sample mean (a) of 3.0 or less? (b) of 4.0 or 5.0 or 6.0? (c) greater than 5.0?

5. Consider a population of five scores: 1, 2, 4, 7, and 11. (a) Find the mean and standard deviation of this set of scores. (b) List the 25 possible samples of size 2 that may be drawn from this population by sampling with replacement. (c) Calculate the mean of each of the 25 samples found in (b). (d) Put the 25 means into a frequency distribution. (e) Find the mean and standard deviation of the distribution formed in (d). (f) Calculate the standard error of the mean of samples of size 2: $\sigma_{\overline{X}} = \sigma/\sqrt{2}$. (g) Does $\mu_{\overline{X}} = \mu_X$? (h) Is $\sigma_{\overline{X}}$ as calculated in (f) the same as the standard deviation calculated in (e)? (i) The population of five scores does not form a symmetrical distribution. Does the distribution of means found in (d) appear to follow the normal curve? If not, does it seem closer to a normal distribution than the distribution formed by the population of five scores? What principle is involved?

6. Given: A normally distributed population with $\mu = 150$ and $\sigma = 24$. If samples of size 36 are drawn at random, what is the probability of obtaining (a) a sample mean of 154 or higher? (b) a sample mean that differs from 150 by 6 points or more?

7. Given: A normally distributed population with $\mu = 120$ and $\sigma = 10$. If samples of size 25 are drawn at random, what is the probability of obtaining (a) a sample mean of 118.25 or lower? (b) a mean that differs from 120 by 2 points or more?

8. Given: A normally distributed population with $\mu = 150$ and $\sigma = 25$. If samples of size 100 are drawn at random, what is the probability of obtaining (a) a sample mean of 156 or higher? (b) a mean that differs from 150 by 7 points or more?

9. For the data of Problem 6, (a) what sample mean is so great that it would be exceeded only 1% of the time in random sampling? (b) Within what central limits would 99% of the sample means fall?

10. For the data of Problem 7, (a) what sample value is so small that the probability of one smaller is .02? (b) Within what central limits would 95% of the sample means fall?

11. For the data of Problem 8, (a) what sample mean is so high that 10% of means would be higher? (b) Within what central limits would 90% of the sample means fall?

12. The Wechsler Adult Intelligence Scale is standardized to have a mean of 100 and a standard deviation of 15. If we draw a random sample of 25 people from the population, what is the probability of obtaining (a) a sample with a mean IQ of 95 or less? (b) a sample with a mean IQ of 105 or greater? (c) Within what central limits would 95% of the sample means fall?

13. For Problem 12, suppose instead that we draw a random sample of 100 people. What is the probability of obtaining (a) a sample with a mean IQ of 95 or less? (b) a sample with a mean IQ of 105 or greater? (c) Within what central limits would 95% of the sample means fall? (d) Compare these answers with those for Problem 12. What general conclusion can you make?

CHAPTER 12

Introduction to Statistical Inference: Testing Hypotheses about Single Means (z and t)

When you have finished studying this chapter, you should be able to:

- Understand that the aim of statistical inference is to form a conclusion about a population parameter from a sample statistic;
- Understand that the key to hypothesis testing is to discover what sample values will occur by chance in repeated sampling and with what probability;
- Formulate the null and alternative hypotheses;
- Know what is meant by level of significance and how it is used to test hypotheses;
- Know when to use a one-tailed or a two-tailed test;
- Know how to test a hypothesis about the population mean when the population standard deviation is known (z) and when it is unknown (t);
- Use IBM SPSS to calculate t;
- Understand the assumptions in testing hypotheses about the mean of a population; and
- Understand the difference between the level of significance and a p-value.

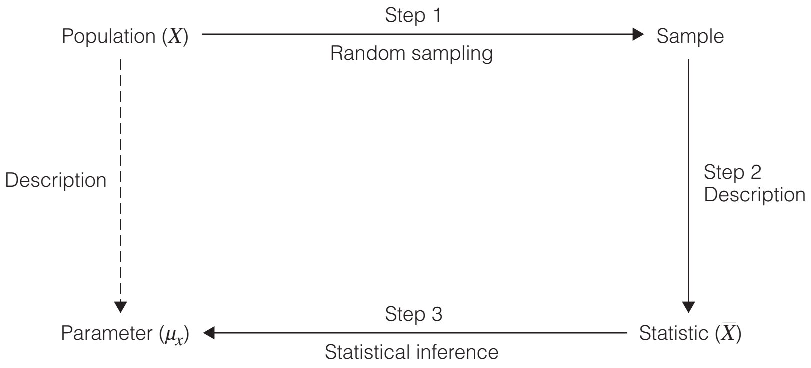


FIGURE 12.1 The steps used in statistical inference (modified from Games & Klare, 1967).

Let us now return to the question asked by Dr. Brown at the beginning of Chapter 11. Suppose that Dr. Brown, the research director of a large school district, has read in a test manual that the national norm for sixth-grade students on a mathematics achievement test is 85. She wants to know whether students in her district are performing at the same level. She selects at random 100 students from all the sixth-grade students in the district and gives the same test to them. Dr. Brown then calculates their mean score and compares it to the national norm. If the mean of her group differs from the national norm, what should she conclude?

statistical inference draws a conclusion about a population parameter from a sample statistic

This is a problem in statistical inference. Recall from Chapter 11 that the basic aim of **statistical inference** is to form a conclusion about a population parameter from a sample taken from that population. That parameter may be a mean, a median, a standard deviation, a correlation coefficient, or any other of a number of statistical parameters. In this chapter, we will concern ourselves with the mean.

Look at Figure 12.1. Ideally, if we wanted to know the value of a population parameter (for example, μ_X), we would gather all the observations in the population and, using the techniques we learned in the chapters on descriptive statistics, directly calculate the parameter. This direct approach is represented by the dashed line. But what if the population is too large for us to take the direct approach? That is when we use statistical inference. Follow the solid lines in Figure 12.1. In Step 1, we randomly draw a sample from the population of interest. In Step 2, we use the same descriptive techniques to calculate the sample statistic (in our example, $\overline{X}$). From the obtained value of our sample statistic ($\overline{X}$), we now infer the value of the population parameter (μ_X).

The population of interest to Dr. Brown consists of the test scores of *all* 2,500 currently enrolled sixth-grade students in her district. From her sample mean ($\overline{X}$), she wants to draw an inference as to whether μ_X, the mean of her population, is 85. However, her sample of 100 student scores is just one of numerous random samples she could have drawn from the population. Suppose, for example, that Dr. Brown replicated her study several times. On each replication, she would randomly draw a different selection of 100 students from the population. Because of sampling variation, we do not expect the means of all the samples to be the same. The value of $\overline{X}$ would probably be different from sample to sample (look again at Section 11.3 and Figure 11.1). How, then, can Dr. Brown infer the value of the population mean when *her* sample mean is only one of many different possible sample means?

12.1 Testing a Hypothesis about a Single Mean

To answer her question (do sixth-grade students in her school district perform as well on the mathematics achievement test as students nationally?), Dr. Brown first translates her question into a statistical hypothesis: $\mu_X = 85$ (the mean of the population of sixth-grade

random sampling distribution of $\bar{X}$
a theoretical relative frequency distribution of all the values of $\bar{X}$ that would be obtained by chance from an infinite number of samples of a particular size drawn from a given population

mathematics achievement scores in her district is 85). To examine the validity of her hypothesis, Dr. Brown asks what sample means would occur by chance if many samples of the same size were drawn at random from her population *if the hypothesis that the population mean is 85 is true.* Dr. Brown now refers to the **random sampling distribution of $\bar{X}$** that you learned about in Chapter 11. She envisions the theoretical random sampling distribution of $\bar{X}$ for an infinite number of samples of size 100 drawn from a population whose mean is 85, and she compares her sample mean with those in this sampling distribution. *The key to any problem in statistical inference is to discover what sample values will occur by chance in repeated sampling and with what probability.*

Suppose the relation between Dr. Brown's sample mean and those of the random sampling distribution of $\bar{X}$ looks like that pictured on the left in Figure 12.2. If so, her sample mean is one that could reasonably occur if the hypothesis is true, and she will retain the hypothesis as a plausible possibility. If the relationship is like that pictured on the right, on the other hand, her sample mean is so far below 85 that it would be quite unusual to obtain such a value by chance when the hypothesis ($\mu_X = 85$) is true. In this case, she will reject the hypothesis and conclude that it is more likely that the mean of her population is not 85.

12.2 The Null and Alternative Hypotheses

null hypothesis
a hypothesis about a population parameter (e.g., μ_X) that a researcher tests

H_0
symbol for the null hypothesis

alternative hypothesis
a hypothesis about a population parameter that contradicts the null hypothesis; in research, the hypothesis the researcher wishes to prove is true

H_A
symbol for the alternative hypothesis

The hypothesis that a researcher tests is called the **null hypothesis**, symbolized **H_0**. It is the hypothesis that he or she will decide to retain or reject. In Dr. Brown's problem, H_0: $\mu_X = 85$. For every null hypothesis, there is also an **alternative hypothesis**, symbolized **H_A**. Because Dr. Brown is interested in discovering a difference (if one exists) *irrespective of direction*, her alternative hypothesis is expressed as H_A: $\mu_X \neq 85$. Notice that *both* H_0 *and* H_A *are always statements about the population parameter.* They are never statements about the sample statistic.

Notice also that the alternative hypothesis is true if the mean mathematics achievement score for *all* sixth graders in Dr. Brown's district is *either* greater than *or* less than 85. Had Dr. Brown been interested in knowing *only* whether mathematics achievement scores in her district are *worse* than the national norm, her alternative hypothesis would have been expressed as H_A: $\mu_X < 85$; or, if *better* than the national norm, as H_A: $\mu_X > 85$. (We will discuss H_A in more detail in Section 12.7.) In either case, if the evidence Dr. Brown obtains in her sample is contrary to H_0 and indicates with sufficient strength that H_A is true, she will reject H_0; otherwise, she will retain it. The decision to reject or retain the hypothesis always refers to H_0 and never to H_A. It is H_0 that is the subject of the test.

Although H_0 is referred to as the null hypothesis, no special meaning should be attached to the term *null*. It does not mean "zero" or "no difference." The null hypothesis is simply whatever hypothesis we choose to test.

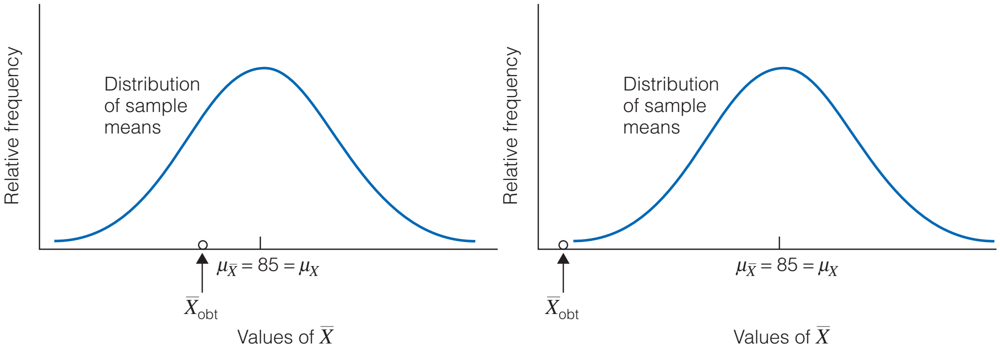
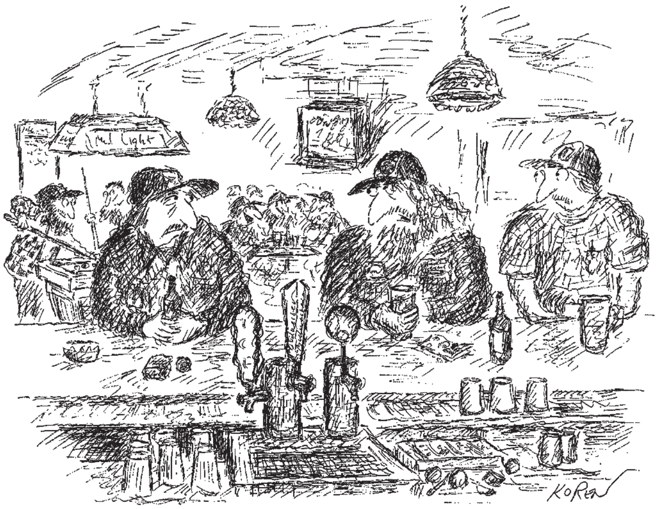


FIGURE 12.2 Possible locations of the obtained sample mean relative to the distribution of expected sample means when the hypothesis is true.



"Are you just pissing and moaning, or can you verify what you're saying with data?"

© The New Yorker Collection, 1999. Edward Koren from cartoonbank .com. All Rights Reserved.

12.3 When Do We Retain and When Do We Reject the Null Hypothesis?

level of significance the probability value that is used as a criterion to decide that an obtained sample statistic ($\overline{X}$) has a low probability of occurring by chance if the null hypothesis is true (resulting in rejection of the null hypothesis)

What criterion should Dr. Brown use to decide when to reject the null hypothesis? When we draw a random sample from a population, *our obtained value of* $\overline{X}$ *will almost never exactly equal* μ_X (because of chance when randomly drawing observations for our sample). Thus, the decision to reject or to retain the null hypothesis depends on the selected criterion for distinguishing between those $\overline{X}$s that would be common and those that would be rare *if* H_0 *is true.*

If the sample mean is so different from what is expected when H_0 is true that its appearance would be unlikely to have occurred by chance, H_0 should be rejected. But what degree of rarity of occurrence is so great that it seems better to reject the null hypothesis than to retain it? This decision is somewhat arbitrary, but common research practice is to reject H_0 if the sample mean is so deviant that its probability of occurrence by chance in random sampling is .05 or less. Such a criterion is called the **level of significance** and is symbolized by the Greek letter **α (alpha)**. In some cases, the researcher may wish to be even more stringent and use a level of significance of .01 or less, but it is rare to choose one greater than .05. We discuss issues governing the choice of level of significance in Section 13.3.

12.4 Review of the Procedure for Hypothesis Testing

α (alpha) symbol for the level of significance

To summarize, here is the procedure for testing *all* statistical hypotheses, whether about means, frequencies, or other population characteristics:

Step 1: A specific hypothesis, called the null hypothesis (H_0), is formulated about a parameter of the population (e.g., μ_X) along with an alternative hypothesis (H_A).

Step 2: A random sample is drawn from the population of observations, and the value of the sample statistic (e.g., $\overline{X}$) is obtained.

Step 3: The random sampling distribution of the statistic under consideration is examined to learn what sample outcomes would occur by chance over an infinite number of repetitions (and with what relative frequencies) if H_0 is true.

Step 4: H_0 is retained if the particular sample outcome is in line with the outcomes expected *if* the hypothesis is true; otherwise, it is rejected and H_A is accepted.

12.5 Dr. Brown's Problem: Conclusion

region of rejection the portion of the sampling distribution of $\overline{X}$ (consisting of values of $\overline{X}$ that are unlikely to have occurred by chance if H_0 is true) that leads to rejection of H_0

region of retention the portion of the sampling distribution of $\overline{X}$ that leads to retention of H_0

In Dr. Brown's problem, H_0: $\mu_X = 85$. Let us assume that Dr. Brown wishes to adopt the .05 level of significance. She therefore will reject H_0 only if her obtained sample mean is so deviant that it falls in the upper 2.5% or lower 2.5% of all the possible sample means that would occur by chance when H_0 is true ($.025 + .025 = .05$). If her $\overline{X}$ falls in the central 95% of the values of $\overline{X}$ that could occur by chance when H_0 is true, she will retain the null hypothesis. The portions of the sampling distribution that include values of $\overline{X}$ that lead to rejection of the null hypothesis (in our example, the two extreme tails) are called the **regions of rejection**. The portion leading to retention of the null hypothesis (in our example, the central 95% of the sampling distribution) is called the **region of retention**.

What sample means would occur by chance if H_0 is true? If it is true, the random sampling distribution of $\overline{X}$ would center on μ_{hyp}, or 85, as shown in Figure 12.3. If Dr. Brown can reasonably assume that the random sampling distribution of $\overline{X}$ approximates a normal curve (and with $n = 100$ she can, regardless of the shape of the population of scores, according to the central limit theorem), then she can use the normal curve table to calculate the z values that separate the upper 2.5% and lower 2.5% of sample means from the remainder. These

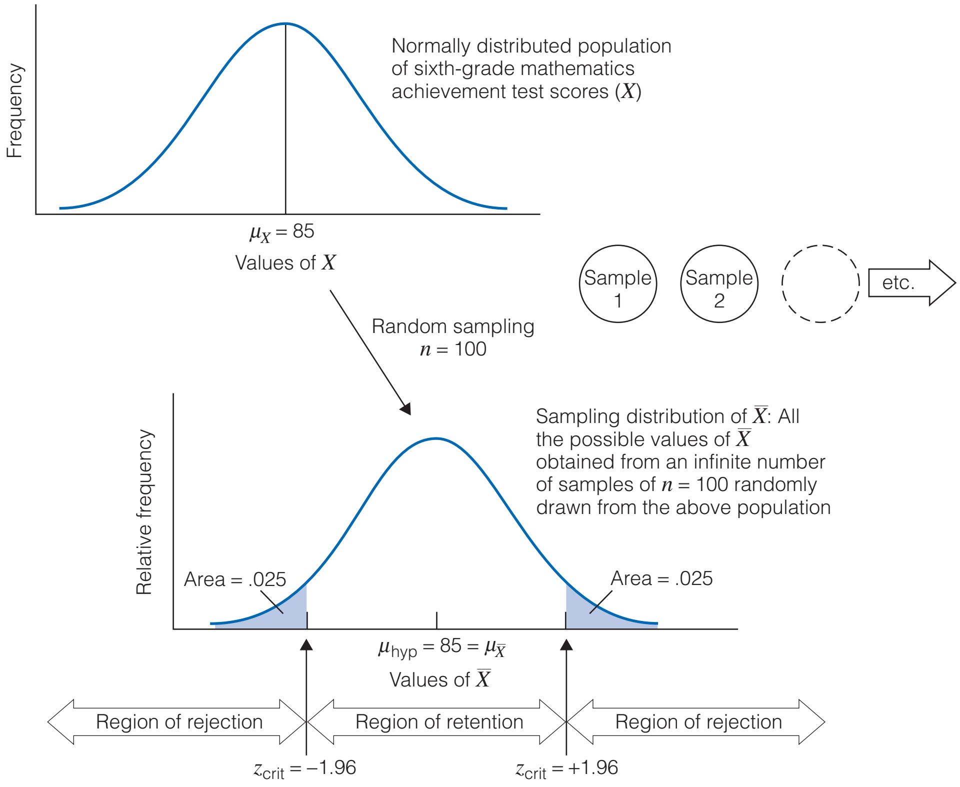

Frequency
Normally distributed population of sixth-grade mathematics achievement test scores (X)
$\mu_X = 85$
Values of X
Sample 1
Sample 2
etc.
Random sampling
$n = 100$
Sampling distribution of $\overline{X}$: All the possible values of $\overline{X}$ obtained from an infinite number of samples of $n = 100$ randomly drawn from the above population
Relative frequency
Area = .025
Area = .025
$\mu_{\text{hyp}} = 85 = \mu_{\overline{X}}$
Values of $\overline{X}$
Region of rejection
Region of retention
Region of rejection
$z_{\text{crit}} = -1.96$
$z_{\text{crit}} = +1.96$

FIGURE 12.3 Dr. Brown's problem: The random sampling distribution of $\overline{X}$ for samples of $n = 100$ drawn at random from a population with $\mu_X = 85$; $\alpha = .05$.

critical value(s) the value(s) that separates the region of rejection from the region of retention

are referred to as the **critical values**. From Table A in Appendix D, the values are found to be: $z = \pm 1.96$.

Suppose Dr. Brown's obtained sample mean is 90. Even if the population mean really is 85, because of random sampling variation we do not expect the mean of a sample randomly drawn from that population to be exactly 85 (although it could be). Thus, the important question is, what is the relative position of the obtained sample mean among all those that could have been obtained by chance *if* the null hypothesis is true? To determine the position of the obtained $\overline{X}$, it must be expressed as a z score:

$$z = \frac{\text{score} - \text{mean of scores}}{\text{standard deviation of scores}}.$$

$$z = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_{\overline{X}}}$$

formula for z required for testing hypotheses about single means

DEFINITIONAL FORMULA FOR z
REQUIRED FOR TESTING HYPOTHESES
ABOUT SINGLE MEANS

$$z = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_X/\sqrt{n}} \tag{12.1}$$

Look at this formula closely. What it says is:

$$z = \frac{\text{obtained sample mean} - \text{hypothesized population mean (when } H_0 \text{ is true)}}{\text{standard error of the mean}}$$

Thus, this z value shows by how many standard errors of the mean the obtained sample mean deviates from the hypothesized population mean stated in H_0. To convert her $\overline{X}$ to z, Dr. Brown must know σ_X. Let us assume that it is 20. So,

$$z = \frac{90 - 85}{20/\sqrt{100}} = \frac{+5}{2} = +2.5$$

In our example, Dr. Brown's sample mean is 2.5 standard errors of the mean greater than expected *if* the null hypothesis is true. You can see the location of this obtained z in the sampling distribution shown in Figure 12.4. As you can see, the value falls in the upper region of rejection, so Dr. Brown rejects H_0 and accepts H_A. In short, she concludes that it is *not* reasonable to believe that the mean of the population from which the sample came is 85, but instead is greater than 85. Notice that *the conclusion is about the population and not simply about the particular sample itself.*

12.6 The Statistical Decision

Has Dr. Brown proven absolutely that the mean of the population is greater than 85? Suppose she had used $\alpha = .01$ as her decision criterion rather than $\alpha = .05$. Her critical values of z that separate the regions of rejection from the region of retention (from Table A in Appendix D) would be different, $z = \pm 2.58$, as shown in Figure 12.5. Dr. Brown's obtained sample mean now falls in the region of retention, and her decision will be to retain H_0. Thus our final decision may depend on our decision criterion. We will return to this in Section 13.3.

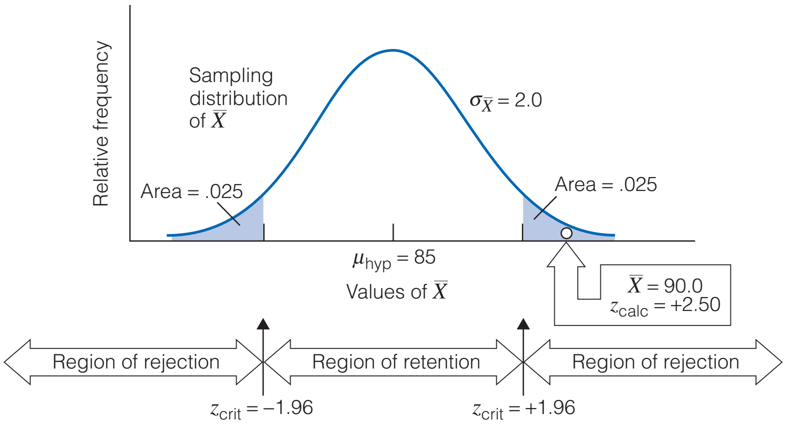


FIGURE 12.4 Dr. Brown's problem: The sampling distribution of $\overline{X}$ for samples drawn at random from a population with $\mu_X = 85$ and $\sigma_X = 20$; $\alpha = .05$. Dr. Brown's sample mean falls in the region of rejection.

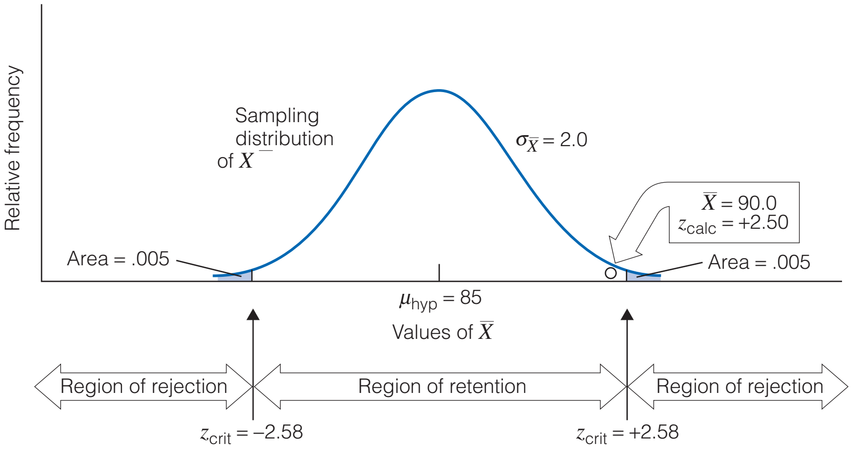


FIGURE 12.5 Dr. Brown's problem: The sampling distribution of $\overline{X}$ for samples drawn at random from a population with $\mu_X = 85$ and $\sigma_X = 20$; $\alpha = .01$. Dr. Brown's sample mean falls in the region of retention.

rejecting the null hypothesis the obtained sample statistic ($\overline{X}$) has a low probability of occurring by chance if the value of the population parameter stated in H_0 is true

retaining the null hypothesis we do not have sufficient evidence to reject H_0 (it does not mean that H_0 must be true, only that it *could* be true)

What does it mean when our sample mean does not fall in the region of rejection and we retain H_0? *The decision to "retain"* H*₀* *does not mean that it is likely that* H*₀* *is true.* Rather, this decision merely reflects the fact that we do not have sufficient evidence to reject the null hypothesis. Other null hypotheses that might have been stated would also have been retained if tested in the same way. Consider our example where the hypothesis is H_0: $\mu_X = 85$. If our sample mean is 86 and is related to the hypothesized sampling distribution of $\overline{X}$ as illustrated at the left in Figure 12.6, our decision will be to retain the null hypothesis. But suppose the hypothesis had been H_0: $\mu_X = 87$. If the same sample result had occurred, $\overline{X} = 86$, we would again be led to retain the null hypothesis that $\mu_X = 87$. This is shown at the right in Figure 12.6.

In short, **rejecting the null hypothesis** means that it does not seem reasonable to believe that it is true, but **retaining the null hypothesis** means only that we believe that the hypothesis *could* be true. It does *not* mean that it must be true, or even that it is probably true.

Look again at the formula for the *z* test. What would be the effect of changing sample size? Suppose that the actual population mean really does differ from the hypothesized mean.

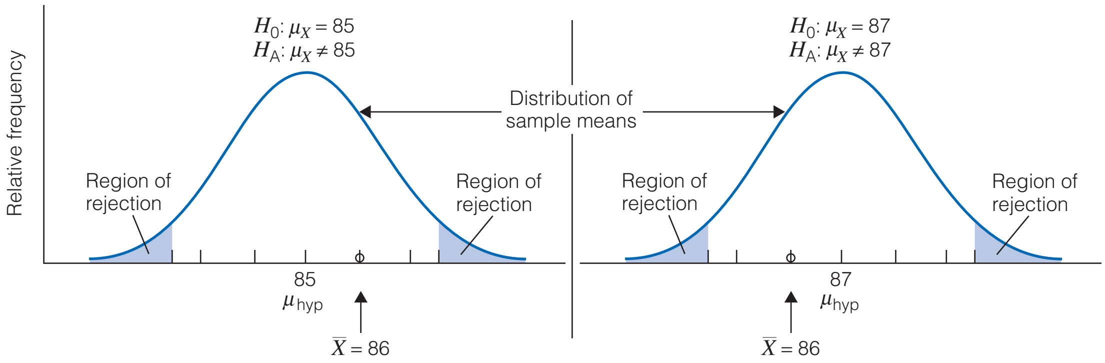


FIGURE 12.6 Testing the null hypothesis that $\mu_X = 85$ or that $\mu_X = 87$ when $\overline{X} = 86$.

If Dr. Brown had drawn a random sample of 25 students instead of 100, she would not have rejected the null hypothesis, because $z = (90 - 85)/(20/\sqrt{25}) = +1.25$, and $+1.25$ falls in the region of retention when $\alpha = .05$.

In general, if the sample we use to test the null hypothesis is small, sampling variation will be large and may mask a difference between μ_{hyp} and μ_{true}. In other words, we might not reject the null hypothesis when we should. Large samples increase precision by reducing sampling variation. However, for a given study there is a limit to the precision needed. In Section 13.11, we will show how to determine optimum size for testing hypotheses about single means.

12.7 Choice of H_A: One-Tailed and Two-Tailed Tests

nondirectional (two-tailed) test the alternative hypothesis, H_A, states that the population parameter may be either less than or greater than the value stated in H_0 (and the critical region is divided between both tails of the sampling distribution)

As you learned in Section 12.2, the alternative hypothesis may be directional or nondirectional, depending on the purpose of the study. When the alternative hypothesis is **nondirectional**, a **two-tailed test** results, and it is possible to detect a difference between the true value and the hypothesized value of the parameter regardless of the direction of the difference. We would want this capability when examining many research questions. For example, in most cases in which the performance of a group is compared with a known standard, it would be of interest to us to discover that the group is superior *or* to discover that the group is substandard.

Sometimes a directional test (a one-tailed test) is more appropriate. Some examples of situations in which a one-tailed test might be suitable are as follows:

1. A social psychologist wishes to know whether children display more aggressive behaviors after watching violence on television.
2. The physical fitness of schoolchildren is tested. If mean performance is substandard, it will be necessary to institute a special physical training program.
3. A new teaching method is proposed and it will be used only if it can be demonstrated that learning accomplished under the new method is superior to that under the standard method.

In a **directional (one-tailed) test**, our interest is in discovering whether or not there is a difference *in a particular direction*. For instance, in Example 2, the question is whether or not the children are substandard with regard to physical fitness. We can test this by hypothesizing that the mean physical fitness score of the population of schoolchildren equals the standard value versus the alternative hypothesis, which is that it is less than this value. Notice that if the children's physical fitness equals or exceeds the standard, no action is necessary. A special training program will be instituted only if their performance is substandard.

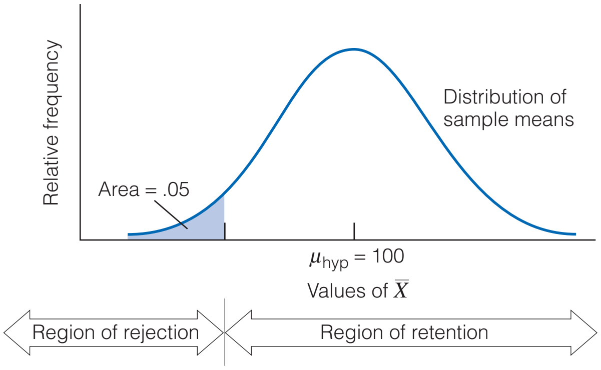


FIGURE 12.7 Testing H_0: $\mu_X = 100$ against H_A: $\mu_X < 100$; $\alpha = .05$.

directional (one-tailed test) the alternative hypothesis, H_A, states that the population parameter differs from the value stated in H_0 in one particular direction (and the critical region is located in only one tail of the sampling distribution)

In a one-tailed test, the alternative hypothesis limits what the researcher has a chance to discover. If H_A says that the population mean is less than μ_{hyp}, then the *only* thing the researcher can discover in this study is that the mean is less than μ_{hyp}. For example, look at Figure 12.7. Here, we are testing H_0: $\mu_X = 100$ against the alternative hypothesis H_A: $\mu_X < 100$. As we can see, the region of rejection is entirely at the left. A value of $\overline{X}$ that falls above 100 *cannot* lead to rejection of the null hypothesis, *no matter how far above 100 it may be.*

The major disadvantage of a one-tailed test is that it does not allow for any chance of discovering that reality is just the opposite of what the alternative hypothesis says. The school board may be interested only in whether a new teaching method is superior (Example 3), but suppose that it actually is worse. They would not conclude this with a one-tailed test where H_A states that the teaching method is greater than normal. If this is unacceptable, a directional alternative is inappropriate and a two-tailed test should be used. In general, *a directional alternative hypothesis is appropriate only when there is no practical difference in meaning between retaining the null hypothesis and concluding that a difference exists in a direction opposite to that stated in the directional alternative hypothesis.*

You may be asking, "Why not wait until you see the results of the statistical test before deciding to use a directional or nondirectional alternative hypothesis?" The decision to use a one-tailed alternative should always flow from the logic of the research question, as just illustrated. Therefore, *the time to decide on the nature of the alternative hypothesis is at the beginning of the study, before the data are collected*. It is not proper to observe the sample outcome and then set the region of rejection in the tail of the sampling distribution toward which the sample outcome tends. For example, if we adopt the 5% level of significance and follow this erroneous procedure systematically, we are really, in the long run, conducting two-tailed tests at the 10% significance level. Likewise, it is not satisfactory to set our one-tailed test in the direction in which we *think* the outcome might go, only to switch to a two-tailed test if the sample mean appears to fall in the opposite direction. If tests are conducted in this manner, using $\alpha = .05$, they are equivalent to two-tailed tests at $\alpha = .075$, with an area of .05 in one tail and an area of .025 in the other—but with the larger area located by the prejudice of preconception. It is because of misuses like these that most journal editors look unkindly at one-tailed tests.

12.8 Review of Assumptions in Testing Hypotheses about a Single Mean

Whenever we perform a z test, we use the normal curve table. For the normal curve model of statistical inference about single means to be precisely correct, several conditions must hold:

1. A random sample has been drawn from the population.
2. The sample has been drawn by the with-replacement sampling plan.

3. The sampling distribution of $\overline{X}$ follows the normal curve.
4. The standard deviation of the population of scores is known.

A truly random sample is often difficult to achieve in practice. Violation of this assumption may affect the mean and standard deviation of the sampling distribution in unpredictable ways (more about that later).

For the model of inference about single means described in this chapter to be strictly correct, it is assumed that sampling is with replacement. Nevertheless, it is common practice to sample without replacement. The consequent error in inference is quite small as long as the sample size is a small fraction of the population size (see Section 11.6).

The third assumption is that the sampling distribution of $\overline{X}$ is normally distributed. As you learned in Section 11.4, this assumption is quite reasonably approximated when the scores in the population are reasonably close to a normal distribution. When the scores in the population are *not* normally distributed, the central limit theorem comes to the rescue when the sample size is 25 or larger (Section 11.4).

The fourth assumption is that we know σ_X, the standard deviation of the population from which our sample is drawn. Look again at Formula 12.1; σ_X appears in the formula for the z test (it is the numerator of the standard error of the mean). But in actual practice, we will rarely know the value of σ_X. Why not? Look at the formula for σ_X:

$$\sigma_X = \sqrt{\frac{\sum (X - \mu_X)^2}{N}}$$

If we know the value of σ_X, we probably already know the value of μ_X. If we know the value of μ_X, why would we want to estimate it from $\overline{X}$? Thus, in most cases we will have to estimate σ_X from the sample data. Let us now examine how this is done.

Point of Controversy

The Single-Subject Research Design

Hypothesis testing, with its reliance on comparisons of means between samples, has been the traditional way of analyzing data in the behavioral sciences for 100 years. In the chapters to come, you will learn about some recent challenges to hypothesis testing, with some statisticians calling for a ban of significance testing.

In actuality, the controversy is not new. One of the original challenges to statistical hypothesis testing came from scientists and practitioners who championed single-subject designs. Rather than obtain a sample of several observations from a population and then worry about whether or not it is representative of that population, why not just use a single subject and study that subject in depth? Early researchers such as Ebbinghaus (human memory) and Pavlov (conditioned responses) used single-subject designs, and this approach is common today in psychophysics and applied behavioral analysis. Perhaps the two strongest proponents of this approach were Sidman (1960) and the famed B. F. Skinner (1966).

Skinner and his colleagues totally rejected comparison of group means and inferential statistics, and eventually created *The Journal of the Experimental Analysis of Behavior*, which was devoted to publishing the results of studies of single subjects. In single-subject experimentation, the dependent variable is measured at multiple times during the study, and the results are not averaged with results from other subjects. The subjects provide their own baselines. In experimental analysis of behavior, for example, the dependent variable is often the subject's response rate, and results are displayed in a real-time graph in which changes in the independent variable across time are shown on the abscissa and changes in the dependent variable are shown on the ordinate. Conclusions about the effects of manipulating the independent variable are made by inspection of the graph (e.g., for changes in response rate after introduction of the variable).

This approach of assessing behavior change is also used to assess the effects of therapy interventions in individual subjects. Proponents claim that collapsing data across subjects obscures "the orderly and systematic development" of behavior in an individual organism (Hopkins, Cole, & Mason, 1998; Morgan & Morgan, 2001; Perone, 1999). In a review, Morgan and Morgan (2001) concluded that

> the single-participant alternative features a design power and flexibility well suited to both basic science and applied research. . . . [Its] features should be especially appealing to practicing clinicians who are delivering services to individual clients and whose professional responsibilities increasingly include documentation of treatment efficacy.

observer bias observations and conclusions that are biased by the observer's own emotions and/or values

Not surprisingly, many researchers oppose the single-subject design. For example, in another review, Jasienski (1996) called it "seriously flawed." There is a risk of **observer bias** (observations and conclusions biased by the observer's own emotions, moods, and values) and there are no control subjects. This is a particular weakness in pharmacological studies, in which one must be alert to the possibility of a *placebo effect* (e.g., the subject feeling better because he or she believes the treatment will work, even if in fact it does not).

The risk of a placebo effect is particularly great in self-experimentation, for which there is a long history in biomedical science (Brown, 1995). According to Jasienski, the major flaw is the lack of individual differences. In a much-publicized case in 1996, baboon bone marrow was used to treat a single AIDS patient (Jasienski, 1996). The treatment failed, but even if it had succeeded, how could the results have been interpreted? The patient had received considerable attention from doctors and a variety of prior drug treatments and radiation. There is also the possibility that the patient was biologically and medically unique in some way. Which aspect of the treatment or subject characteristic, or combination of factors, was responsible for the outcome? In every measurement there is also some degree of error.

Proponents of hypothesis testing say that the only way to deal with these problems is by randomly drawing samples from a population and treating everyone the same except for the independent variable. There will still be individual differences and errors in measurement (however small), but "without statistics, there is no way of knowing if the errors matter. . . . The only way is through 'averaging out' the superficial differences among individuals, to see what meaningful similarities remain" (Jasienski, 1996). Even Morgan and Morgan (2001) admit that "when . . . one is interested in population parameters for the purpose of establishing social policies or regulations affecting educational, political, or social institutions, then group designs may have considerable usefulness."

Although single-subject design research without formal inferential statistics still has its defenders (e.g., Hopkins, Cole, & Mason, 1998; Morgan & Morgan, 2001; Perone, 1999), there is a growing awareness of the use of inferential statistics in such designs (e.g., Franklin, Allison, & Gorman, 1997) and even manuscripts submitted to *The Journal of the Experimental Analysis of Behavior* have increasingly included inferential statistics (Ator, 1999).

12.9 Estimating the Standard Error of the Mean When σ Is Unknown

When we test hypotheses about means using z, we must calculate the standard error of the mean ($\sigma_{\overline{X}}$). In Dr. Brown's problem, we gave its value. However, the standard error of the mean requires that we know σ, the standard deviation of the population, which in actual practice we must almost always estimate from the sample:

$$z = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_X/\sqrt{n}} = \frac{\overline{X} - \mu_{\text{hyp}}}{?/\sqrt{n}}$$

unbiased estimator the mean of the estimates made from all possible samples equals the value of the parameter estimated ($\overline{X}$ is an unbiased estimator of μ_X; S_X is not an unbiased estimator of σ_X)

Intuition suggests that we substitute S_X, the sample standard deviation, as an estimate of σ_X, but as it turns out, this creates an additional problem.

The basic problem is that the sample variance, S_X^2, is a *biased estimator* of the population variance, σ_X^2. What we need is an unbiased estimator. With an **unbiased estimator**, we do not expect our sample value to exactly equal the population parameter, but the mean of the estimates made from all possible samples equals the value of the parameter estimated. For example, the sample mean is an unbiased estimator of the population mean because the average of the values of $\overline{X}$ computed from all possible samples exactly equals μ_X. However, the mean value of S_X^2 calculated from all possible samples is a little smaller than σ_X^2. We introduced this problem, as well as the following correction, in the Point of Controversy in Chapter 5.

Recall that the formula for the sample variance (Formula 5.2*b*) is:

$$S_X^2 = \frac{\sum(X - \overline{X})^2}{n}$$

It has been shown that the tendency toward underestimation will be corrected if $\Sigma(X - \overline{X})^2$ is divided by $n - 1$ rather than by n. The following formula incorporates this correction:

SS_X symbol for the sum of the squared deviations from the mean

UNBIASED ESTIMATE OF THE
POPULATION VARIANCE

$$s_X^2 = \frac{\sum(X - \overline{X})^2}{n - 1} = \frac{SS_X}{n - 1} \qquad (12.2)$$

$\boldsymbol{s_X^2 = \frac{SS_X}{n - 1}}$ unbiased estimate of the population variance

The use of the lowercase letter in s_X^2 distinguishes this formula from the one for S_X^2. Taking the square root of the formula, we obtain a better estimate of the standard deviation of the population:

$$s_X = \sqrt{\frac{\sum(X - \overline{X})^2}{n - 1}}$$

$\boldsymbol{s_X = \sqrt{\frac{SS_X}{n - 1}}}$ a better estimate (than S_X) of σ_X

We give a mathematical explanation of why S_X^2 is a biased estimator of σ_X^2 in Note 12.1 at the end of the chapter. However, an empirical demonstration may be easier to understand. Look at Table 12.1. It shows the 16 different samples that are possible when we draw samples of size 2 from a population of four observations. When we calculate the variance of the four scores by Formula 5.2*a*, we get $\sigma_X^2 = 5.0$. Now let us calculate s_X^2 using Formula 12.2 from each of the 16 possible samples. The 16 samples and the 16 variance estimates are shown, respectively, in the first and last columns of the table.

estimated standard error of the mean an estimate of the standard deviation of the random sampling distribution of $\overline{X}$

Notice that the mean of these estimates is 5.0, which is also the true value of σ_X^2, the quantity estimated. This demonstrates that s^2 is an unbiased estimator of σ^2 because the mean of all the estimates equals the quantity estimated. Obviously, if we had made the 16 estimates by dividing by $n = 2$ rather than $n - 1 = 1$, the mean of *these* estimates would have been less than 5.0. Although the correction for estimating the standard error of the mean results in a better estimate *on the average*, recognize that any particular sample will probably result in an estimate that is either too large or too small.

$\boldsymbol{s_{\overline{X}} = s_X/\sqrt{n}}$ symbol and formula for the estimated standard error of the mean ($\overline{X}$)

When we substitute s_X for σ_X, the result is called the **estimated standard error of the mean**, symbolized $s_{\overline{X}}$:

TABLE 12.1 *Possible Samples, Together with Estimates of the Population Variance Computed for Each (Data from Table 11.1)*

POPULATION: 2, 4, 6, 8

$$\textbf{POPULATION VARIANCE: } \sigma_X^2 = \frac{\sum(X-\mu_X)^2}{N} = 5.0$$

Sample	$\overline{X}$	$\sum(X-\overline{X})^2$	*Estimate of population variance* $s_X^2 = \dfrac{\sum(X-\overline{X})^2}{n-1}$
2,2	2	0	0
2,4	3	2	2
2,6	4	8	8
2,8	5	18	18
4,2	3	2	2
4,4	4	0	0
4,6	5	2	2
4,8	6	8	8
6,2	4	8	8
6,4	5	2	2
6,6	6	0	0
6,8	7	2	2
8,2	5	18	18
8,4	6	8	8
8,6	7	2	2
8,8	8	0	0
			Mean value of s_X^2: 5.0

ESTIMATED STANDARD
ERROR OF THE MEAN

$$s_{\overline{X}} = s_X/\sqrt{n} \qquad (12.3)$$

Note that the lowercase *s* signifies an estimate of a standard deviation in both symbols. When we substitute $s_{\overline{X}}$ for $\sigma_{\overline{X}}$ in the formula for *z*, the result is a new statistic (an approximation of *z*) called ***t***. Unlike *z*, however, *t* is not normally distributed.

12.10 The *t* Distribution

If the assumptions of Section 12.8 are satisfied, the value of *z* corresponding to the sample mean is a normally distributed variable. Why? The values of $\overline{X}$ will vary as we move from random sample to random sample, but the values of μ_X and $\sigma_{\overline{X}}$ are constants and will remain fixed. Consequently, the values of *z* are formed as follows:

$$z = \frac{(\textit{normally distributed variable}) - (\textit{constant})}{(\textit{constant})}$$

where: $\overline{X}$ = the normally distributed variable
μ_X = a constant
$\sigma_{\overline{X}}$ = a constant.

Subtracting a constant from each score in a normal distribution does not change the shape of the distribution, nor does dividing by a constant (see Section 5.10). Consequently, z will be normally distributed when $\overline{X}$ is normally distributed, and the normal distribution is therefore the correct distribution to which to refer z for evaluation.

Consider the same problem, but this time assume σ is not known and must be estimated from the sample. To evaluate the position of the sample mean, we calculate the statistic t:

$$t = \frac{\overline{X} - \mu_X}{s_{\overline{X}}}$$

formula for the t statistic; used when the population standard deviation (σ_X) is not known

DEFINITIONAL FORMULA FOR t REQUIRED FOR TESTING HYPOTHESES ABOUT SINGLE MEANS

$$t = \frac{\overline{X} - \mu_X}{s_{\overline{X}}} \tag{12.4a}$$

The value μ_X is the value stated in the null hypothesis (μ_{hyp}). So,

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}} \tag{12.4b}$$

Now look again at Formula 12.1, the formula for z that is required for testing hypotheses about single means. We may summarize the formula for t as

$$t = \frac{\textit{obtained sample mean} - \textit{hypothesized population mean (when } H_0 \textit{ is true)}}{\textit{estimated standard error of the mean}}$$

Recall from Formula 12.3 that $s_{\overline{X}} = s_X/\sqrt{n}$. The t statistic differs from z in that *the denominator of the expression is a variable.* As we move from random sample to random sample, not only will values of $\overline{X}$ vary but *each sample will probably result in a different estimate* ($s_{\overline{X}}$) of $\sigma_{\overline{X}}$.[1] The resulting statistic is formed as follows:

$$t = \frac{\textit{(normally distributed variable)} - \textit{(constant)}}{\textit{(variable)}}$$

Student's distribution of *t*
a theoretical relative frequency distribution of all the values of $\overline{X}$ converted to t that would be obtained by chance from an infinite number of samples of a particular size drawn from a given population

If we were to draw an infinite number of samples of the same size n from a population, calculate $\overline{X}$, s_X, and $s_{\overline{X}}$ for each sample, and then find t for each sample, we would have the sampling distribution of t. Because of the presence of the variable in the denominator, this statistic does *not* follow the normal distribution. Just what distribution does it follow? Shortly after the turn of the century, British mathematician William S. Gosset, writing under the name of "Student" while working at Guinness Brewing Company, solved this problem and presented the proper distribution for it. Since that time, the statistic has been called t, and its distribution has been referred to as *Student's distribution* or **Student's distribution of *t***.

Do not make the error of thinking that it is the sampling distribution of $\overline{X}$ that becomes nonnormal when σ_X must be estimated from the sample. If the assumptions of Section 12.8 are met, $\overline{X}$ will be normally distributed regardless of the sample size. However, the position of $\overline{X}$ is not evaluated directly; rather, it is the value of the statistic $t = (\overline{X} - \mu_X)/s_{\overline{X}}$. *Although* $\overline{X}$ *is normally distributed, the resulting value of* t *is not,* for the reason cited previously.

---

[1] The sampling distribution of $s_{\overline{X}}$ tends to be positively skewed.

12.11 Characteristics of Student's Distribution of *t*

What are the characteristics of Student's sampling distribution of *t*? When samples are large, the values of $s_{\overline{X}}$ will be close to that of $\sigma_{\overline{X}}$, and the distribution of *t* will be very nearly normal (with an infinitely large sample, $t = z$). On the other hand, when sample size is small, the values of $s_{\overline{X}}$ will vary considerably from $\sigma_{\overline{X}}$. The distribution of *t* will then also depart considerably from that of the normally distributed *z*.

degrees of freedom (*df*) the number of observations that are completely free to vary; for the variance or standard deviation, $df = n - 1$ (see Section 12.12)

Student's distribution of t *is not a single distribution, but rather a family of distributions.* They differ in their degree of approximation to the normal curve. The exact shape of a particular member of the family depends on the sample size or, more accurately, on the number of **degrees of freedom (*df*)**, a quantity closely related to sample size. For problems of inference about single means, $df = n - 1$. We will explain the concept of degrees of freedom in greater detail in the next section.

The assumptions for Student's distribution of *t* are the same as for the normal curve model (Section 12.8), except that we do not know the population standard deviation. Thus, when the number of degrees of freedom is less than infinite, the theoretical distribution of *t* and the normal distribution *z* are alike in some ways and different in others. They are alike in that both distributions

1. Have a mean of zero.
2. Are symmetrical.
3. Are unimodal.

The two distributions differ in that the distribution of *t*

1. Is *platykurtic* compared to the normal distribution (i.e., it is narrower at the peak and has a greater concentration in the tails than does a normal curve).
2. Has a larger standard deviation (remember that $\sigma_Z = 1$).
3. Depends on the number of degrees of freedom. Figure 12.8 illustrates this last point. It shows the normalized *z* distribution, and, in contrast, Student's distribution of *t* for $df = 4$, $df = 12$, and $df = \infty$.

The differences between the *t* and normal *z* distributions have an important consequence in statistical inference. For example, if we test a null hypothesis at $\alpha = .05$, the critical value of *z* is ±1.96. Because the tails of the *t* distribution descend less rapidly to the horizontal axis, it is necessary to go to a more extreme location to find the point beyond which .025 of the area (in one tail) falls, as Figure 12.8 shows. How much larger than *z* must *t* be to correspond to the

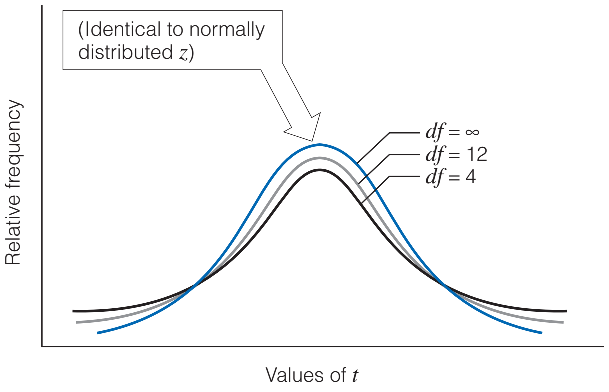


FIGURE 12.8 The distribution of Student's *t* for three levels of degrees of freedom.

TABLE 12.2 *Critical Values for Student's Distribution of* t *for* $\alpha = .05$ *and Selected Degrees of Freedom*

df	5	10	25	50	100	500	∞
$t_{.05}$	±2.571	±2.228	±2.060	±2.008	±1.984	±1.965	±1.960

same rarity of occurrence? It depends on the sample size (or, more accurately, on the number of degrees of freedom). Look at Table 12.2. It shows, for selected values of *df*, the critical magnitude of t corresponding to a two-tailed test conducted according to the decision criterion $\alpha = .05$. Notice that

1. For an infinite number of degrees of freedom, the critical magnitude of t is the same as the critical magnitude of z: $+1.96$.
2. The smaller the number of degrees of freedom, the larger the critical magnitude of t.
3. As the number of degrees of freedom decreases, the change in Student's distribution of t, slight at first, progresses at an increasingly rapid rate.

12.12 Degrees of Freedom and Student's Distribution of *t*

The degrees of freedom associated with s, the estimated standard deviation of a population, corresponds to the number of observations that are completely free to vary. Suppose a sample consists of three scores: X_1, X_2, and X_3. If we calculate s_X by the deviation-score formula, $s_X = \sqrt{\Sigma(X - \overline{X})^2/(n-1)}$, the value of s_X will be determined by the values of the three deviation scores: $(X_1 - \overline{X})$, $(X_2 - \overline{X})$, and $(X_3 - \overline{X})$. Now only two (any two) of the deviation scores are free to vary. We could, if we wish, arbitrarily assign to $(X_1 - \overline{X})$ the value $+3$ and to $(X_2 - \overline{X})$ the value -7. But as soon as two of the three have been given numerical values, the third is no longer free to vary. Do you remember from Chapter 4 that *the sum of deviations taken about the mean is always zero*? (See Section 4.5 and Note 4.1.) In other words, $\Sigma(X - \overline{X}) = 0$. So as soon as the first two deviations are assigned, the third must be a value such that the three deviations sum to zero. In our numerical example, the value of $(X_3 - \overline{X})$ must be $+4$.

For a variance estimate made from a single sample, there will be n deviations from the mean. All but one of them will be free to vary; the last will be determined by the constraint that $\Sigma(X - \overline{X}) = 0$. The number of degrees of freedom in a problem involving the calculation of s is therefore $n - 1$.

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}}$$

formula for t required for testing a null hypothesis about a single population mean

The theoretical distribution of t appears in Table D in Appendix D. The table makes no distinction between negative and positive values of t because the area falling above a given positive value of t is the same as the area falling below the same negative value.

When testing hypotheses about the mean, you find the critical value of t in Table D at the intersection of the row corresponding to the appropriate degrees of freedom (given at the left of the table) and the column corresponding to the appropriate level of significance *for each tail* (given at the top of the table). For example, when $df = 20$ and $\alpha = .05$, what values of t differentiate the regions of rejection from the region of retention? Assume that we wish to discover a difference (if one exists) irrespective of direction. We must identify the values of t such that the central 95% of the obtained ts will be included within these limits. These are (1) the value of t below which 2.5% of the ts will be located and (2) the value of t above which 2.5% of the ts will be located. The required magnitude of t is found under the heading *Area in one tail* = .025. For $df = 20$ the value is $t = \pm 2.086$.

12.13 An Example: Has the Violent Content of Television Programs Increased?

Social psychologists, sociologists, communication researchers, and many others are concerned about the effects on aggression of violent content in the media, particularly television. There is, in fact, a positive correlation between the amount of violence that children and adolescents see on television and aggressive acts that they display towards others (e.g., see Anderson et al., 2003; Christensen & Wood, 2006; Johnson et al., 2002).

Children average about 4 hours a day watching television, with one-fourth of them spending more than 6 hours a day. In the late 1980s, prime-time television shows had three to four violent acts per hour. Thus, at that time, the average child watched about 100 violent acts on television every week. Let us suppose that Professor Telley believes that the media glorifies violence and as a result the amount of violent content that children see on prime-time television shows has increased even further since the late 1980s. To test her belief, Professor Telley draws a random sample of 25 eighth-grade school children from the local school district and calculates the number of violent acts that each of them had seen on television in the previous week.

Her research question leads to a statistical hypothesis to be tested, she decides, at the .05 level of significance:

$$H_0\colon \mu_X = 100$$

Although Professor Telley believes that children are seeing more violent acts on television than in the 1980s, she also wants to know whether they are seeing fewer violent acts. Thus, her alternative hypothesis is expressed as

$$H_A\colon \mu_X \neq 100$$

The mean number of violent acts seen on television in 1 week by her sample of 25 children turns out to be 120. Would drawing a sample with this or a more extreme mean be likely if the null hypothesis is true?

To answer this question, she must estimate the standard error of the mean from her sample data and convert $\overline{X}$ to a t value. Let us assume that s_X proves to be 35.5. Thus,

$$s_{\overline{X}} = \frac{s_X}{\sqrt{n}} = \frac{35.5}{\sqrt{25}} = 7.1$$

and

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}} = \frac{120 - 100}{7.1} = +2.82$$

What does a t value of $+2.82$ indicate? It means that Professor Telley's sample mean was 2.82 estimated standard error of the mean units greater than the value stated in H_0. The number of degrees of freedom is $n - 1 = 24$. For $\alpha = .05$, the critical value of t is the one that in each tail of the distribution divides the outer 2.5% of the area from the remainder. From Table D in Appendix D, we see that the critical value is ± 2.064. In other words, if the null hypothesis is true (that is, $\mu_X = 100$), only 5% of all possible sample means would differ by as much as 2.064 estimated standard errors of the mean. Figure 12.9 shows the sampling distribution of $\overline{X}$ (for when H_0 is true), the region of rejection, and the location of the calculated value of t.

Professor Telley's sample data lead her to reject the null hypothesis and accept the alternative. However, this is just the statistical decision (see Section 1.3). Based on this, Professor

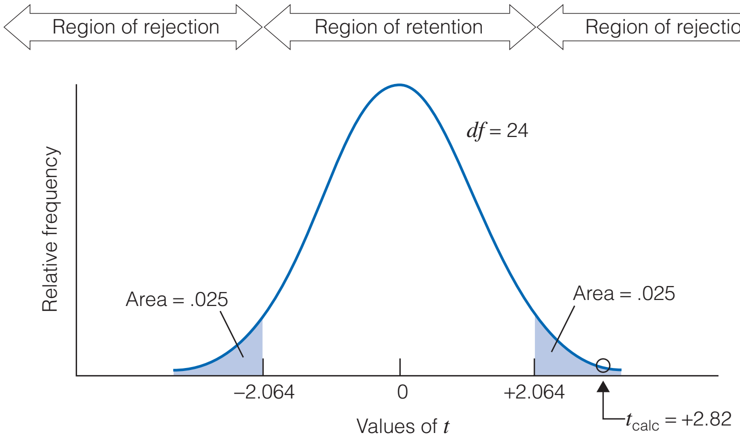


FIGURE 12.9 Testing a hypothesis about μ_X against a nondirectional alternative; $\alpha = .05$.

Telley can now reach a research conclusion. She concludes that eighth-grade children in her local school district see more violent acts on television during prime time than children did 20 years earlier because television and the media glorify violence.

Professor Telley cannot be certain that eighth-grade children in her school district see more violent acts on prime-time television than children in the 1980s. She would have to survey all the eighth-grade school children in her school district to be 100% certain. Inferential statistics, you recall, deals only with probabilities. What inferential statistics says here is that a sample mean of 120 was *unlikely* to occur by chance in random sampling from a population with a mean of 100.

statistical conclusion a conclusion about the numerical property of the data (reject or retain H_0)

research conclusion a conclusion about the subject matter

However, even if Professor Telley's statistical conclusion is correct (let us assume that the mean number of violent acts seen by eighth-grade children in her school district really is greater than the mean number seen in the 1980s), it does not mean that her research conclusion is correct. She might have concluded that the greater number of violent acts was due to the media's glorification of violence, when in fact the television shows that the children watched portrayed the violence as wrong and immoral. Be sure to review Section 1.3 to appreciate the difference between a **statistical conclusion** and a **research conclusion**.

Recall from Section 12.7 that sometimes a directional alternative hypothesis is better suited to the problem at hand. If Professor Telley had been concerned *only* with the possibility that eighth-grade students see more violent acts on television than in the past, she might have adopted the alternative hypothesis H_A: $\mu_X > 100$ and used a one-tailed test. She would have placed the entire region of rejection in the *right* tail of the sampling distribution rather than dividing it between the two tails. Then, she would have rejected H_0: $\mu_X = 100$ only in the face of evidence that μ_X is greater than 100. Figure 12.10 shows how the test would have been conducted. If $\alpha = .05$, the critical value of t is +1.711. Professor Telley's calculated value of t (+2.82) falls in the region of rejection, so her decision would be to reject H_0.

Suppose, however, that Professor Telley had been interested in knowing only whether the eighth-grade children had seen fewer acts of violence than in the past. In that case, she would have adopted the alternative hypothesis H_A: $\mu_X < 100$, and the region of rejection would have been placed entirely in the *left* tail of the distribution. This is illustrated in Figure 12.11. In this event, her sample mean does not fall in the region of rejection so her decision would be to retain H_0. Remember, *the choice between a nondirectional or directional alternative hypothesis should be determined by the rationale that gives rise to the study and should be made before the data are gathered.*

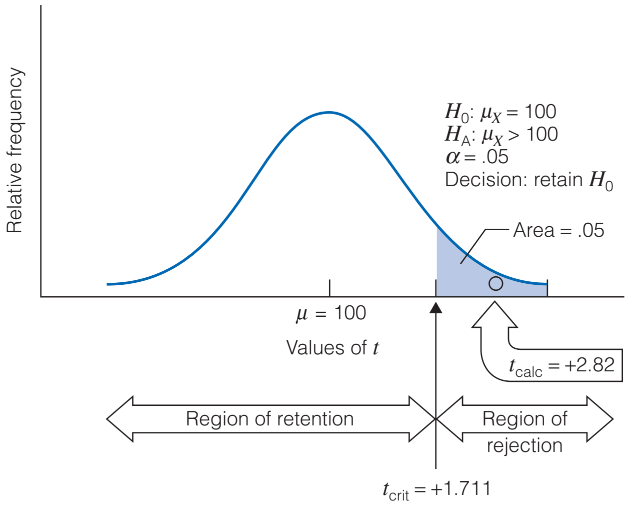

Relative frequency
H_0: μ_X = 100
H_A: μ_X > 100
α = .05
Decision: retain H_0
Area = .05
μ = 100
Values of t
t_{calc} = +2.82
Region of retention
Region of rejection
t_{crit} = +1.711

FIGURE 12.10 Location of regions of retention and rejection when H_A is $\mu_X > 100$; $\alpha = .05$.

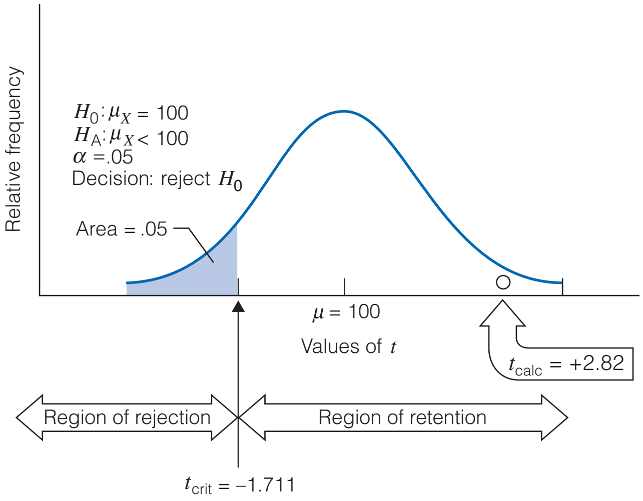

Relative frequency
H_0: μ_X = 100
H_A: μ_X < 100
α =.05
Decision: reject H_0
Area = .05
μ = 100
Values of t
t_{calc} = +2.82
Region of rejection
Region of retention
t_{crit} = −1.711

FIGURE 12.11 Location of regions of retention and rejection when H_A is $\mu_X < 100$; $\alpha = .05$.

12.14 Calculating *t* from Raw Scores

In the real world, we are not given s_X; we must calculate it from our sample data. Of course, computers can perform all of these operations for you (see next section), but there may be occasions when you do not have access to a computer and you have to calculate *t* by hand. A computational formula for computing *t* from raw sample data can easily be derived from Formula 12.4 as follows: From Section 12.10 and Formula 12.4,

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}}$$

From Section 12.9,

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_X/\sqrt{n}}$$

Therefore,

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{SS_X}{n-1}}\Big/\sqrt{n}} = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{\sum(X-\overline{X})^2}{n-1}}\Big/\sqrt{n}}$$

In Section 5.5, you learned that the raw-score method (Formula 5.4) is generally easier to use when calculating the sum of squares (*SS*). Thus,

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{SS_X}{n-1}} \Big/ \sqrt{n}} = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}}{n-1}} \Bigg/ \sqrt{n}}$$

Because both parts of the denominator are under a square root sign, we can express all of it as one square root:

$$\boldsymbol{t = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{SS_X}{n(n-1)}}}}$$

computational formula for *t* for testing a null hypothesis about a single mean

COMPUTATIONAL FORMULA FOR *t* FOR TESTING HYPOTHESES ABOUT SINGLE MEANS

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{SS_X}{n(n-1)}}} \qquad (12.5)$$

Do not lose sight of the fact that the denominator in Formula 12.5 is the same as that in Formula 12.3; both are the estimated standard error of the mean.

We will now show how to use Formula 12.5 with an example. Suppose that Dr. Jones, a colleague of Dr. Telley, wants to replicate her study with a sample of 10th-grade students. The statistical hypothesis is again:

$$H_0\colon \mu_X = 100$$

$$H_A\colon \mu_X \neq 100$$

$$\alpha = .05$$

Dr. Jones now randomly selects fifteen 10th-graders from the same school district used by Dr. Telley and determines how many violent acts each had seen on television in the previous week. The hypothetical data appear in Table 12.3 (①). The mean of the 15 measurements proves to be 110.8 (②). Because the population standard deviation is not available, we must estimate the standard error of the mean according to Formula 12.3 (③). It proves to be 3.055. Our calculated value of t is $+3.535$ (④). This means that Dr. Jones' sample mean of 110.8 is 3.535 estimated standard error of the mean units greater than 100, the value stated in the null hypothesis. Is this a value that would be likely to have occurred by chance if the null hypothesis is true?

Our estimate of the standard error of the mean, $s_{\overline{X}}$, was based on 15 observations, so $df = n - 1 = 14$. For $\alpha = .05$ (2.5% in each tail), the critical value of t (from Table D in Appendix D) is $t = \pm 2.145$. Figure 12.12 shows the critical value of t, the region of rejection, and the location of the calculated value of t. As you can see, this value has a low probability of occurring by chance ($<$2.5%) if the null hypothesis is true. Thus, Dr. Jones rejects the null hypothesis and concludes that 10th graders in her school district see more violent acts on television today than 10th graders in the past.

Did you notice that the means for Dr. Telley's study and Dr. Jones' study differed not only from the value stated in the null hypothesis (100), but also from each other? Which is correct, or is it the case that eighth graders see more violent acts on television than 10th graders?

TABLE 12.3 *Calculation of* t *from the Computational Formula*

① ACTS OF VIOLENCE (X)	X^2
109	11,881
130	16,900
95	9,025
105	11,025
120	14,400
112	12,544
124	15,376
108	11,664
116	13,456
98	9,604
122	14,884
109	11,881
118	13,924
111	12,321
85	7,225
$\sum X = 1662$	$\sum X^2 = 186{,}110$

② $\overline{X} = 110.8$

③ Calculation of $s_{\overline{X}}$:

$$s_{\overline{X}} = \sqrt{\frac{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}{n(n-1)}}$$

$$= \sqrt{\frac{186{,}110 - \frac{(1662)^2}{15}}{15(14)}}$$

$$= \sqrt{\frac{186{,}110 - 184{,}149.6}{210}}$$

$$= \sqrt{9.335}$$

$$= 3.055$$

Calculation of t:

④ $$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}} = \frac{110.8 - 100}{3.055} = +3.535$$

Remember, these are sample means, not means of the population(s). If Dr. Telley drew another sample of 25 eighth graders, it is likely that her sample mean would differ not only from 100 (H_0) again, but also from 120, the mean of her first sample (look at Figs. 11.1 and 12.3 again). The same is true for Dr. Jones. Even if the null hypothesis is incorrect in both studies, it is unlikely that the mean for the population of eighth graders is exactly 120 and that for the population of 10th graders is exactly 110.8. All we can conclude at this point is that in both cases the population means are greater than 100. As for the question of whether the difference in

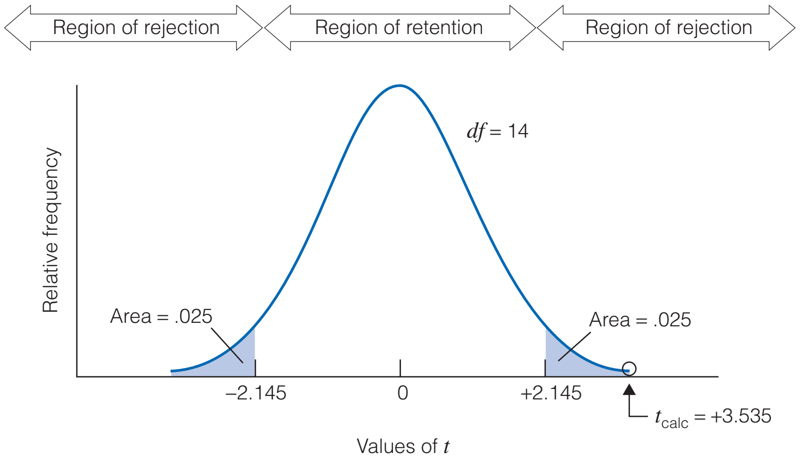


FIGURE 12.12 Testing a hypothesis about μ_X against a nondirectional alternative; $\alpha = .05$.

sample means between eighth graders and 10th graders is large enough to indicate a true difference (as opposed to being due solely to chance), you will learn how to make this type of decision in Chapter 14.

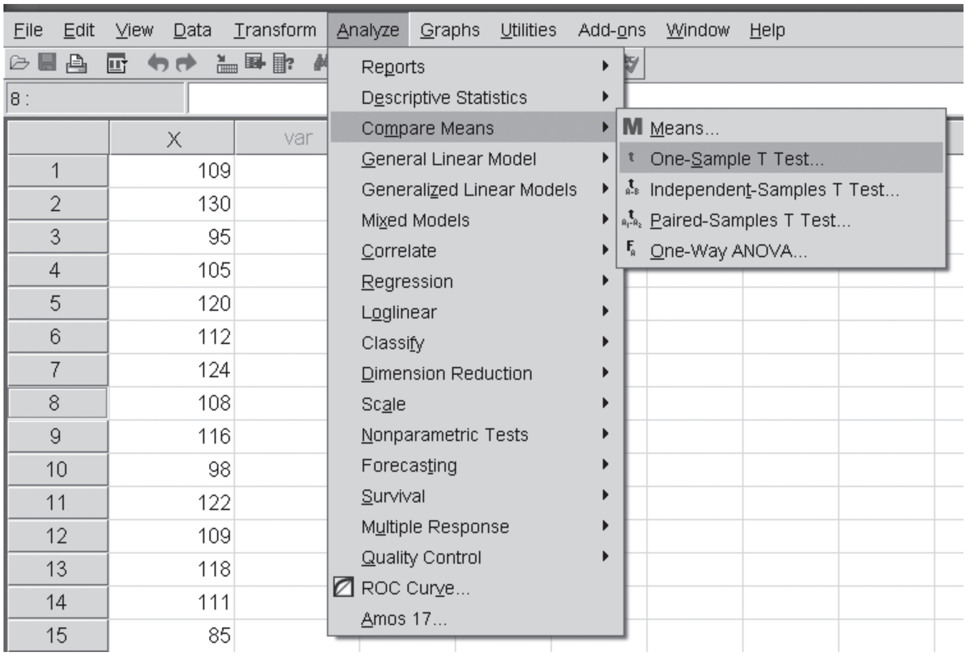
12.15 Calculating *t* with IBM SPSS

In this section, you will learn how to use IBM SPSS to calculate the *t* statistic to test hypotheses about a single mean. You will use the data in Table 12.3. We will assume that you have already entered your data into one column in IBM SPSS's **Data Editor**. Recall from Section 7.8 that you can enter variable names and labels when in the **Variable View** tab. Here, we used X as the variable name and Acts of Violence as the label.

To calculate a one-sample *t* test in IBM SPSS, click **Analyze,** then select **Compare Means** followed by **One-Sample T Test**. Figure 12.13 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will notice the menus that would appear after clicking **Analyze** → **Compare Means** → **One-Sample T Test**.

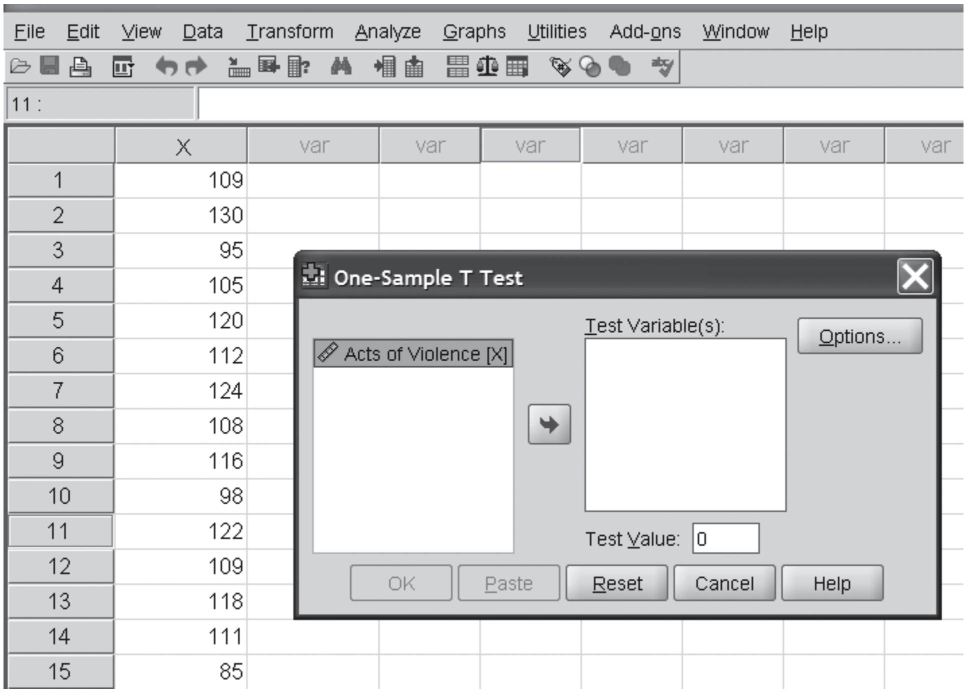
Now, a dialog box will appear like that shown in Figure 12.14. You should move the Acts of Violence variable (i.e., X) to the box labeled Test Variable(s). Next, in the box labeled Test Value, type in the hypothesized value of 100. Figure 12.15 shows a screenshot after the X variable has been moved and the hypothesized value of 100 has been entered. Now, click the OK button.

Table 12.4 shows the IBM SPSS output. After the syntax, there is a section of output labeled One-Sample Statistics. It shows that *n* is 15 and the sample mean is 110.8. In addition, IBM SPSS provides 11.833 as an estimate of the standard deviation in the population (see Formula 12.2) and 3.055 as an estimate of the standard error of the mean (see Formula 12.3). Next, in the section of the output labeled One-Sample Test, the calculated *t* statistic can be found under the column *t*. This value matches the +3.535 in Table 12.3. The next column provides the degrees of freedom. The following column labeled Sig. (2-tailed) will be discussed in Section 12.16. The next column is the value from the numerator of the *t* statistic, in this case, 110.8 − 100 = 10.8. The remaining columns will be discussed in Section 17.2.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
8 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...

	X	var
1	109	
2	130	
3	95	
4	105	
5	120	
6	112	
7	124	
8	108	
9	116	
10	98	
11	122	
12	109	
13	118	
14	111	
15	85	

FIGURE 12.13 IBM SPSS menus when calculating a one-sample *t*.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
11 :
var
One-Sample T Test
Options...
OK
Paste
Reset
Cancel
Help

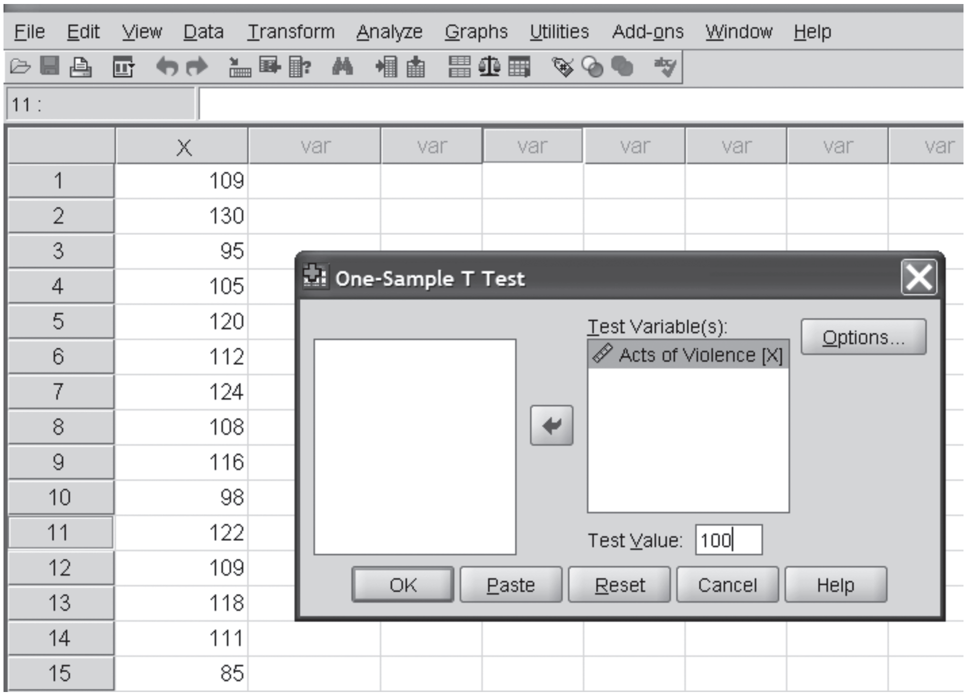
	X
1	109
2	130
3	95
4	105
5	120
6	112
7	124
8	108
9	116
10	98
11	122
12	109
13	118
14	111
15	85

Acts of Violence [X]

Test Variable(s):

Test Value: 0

FIGURE 12.14 IBM SPSS dialog box when calculating a one-sample *t*.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
11 :
var
One-Sample T Test
Options...
OK
Paste
Reset
Cancel
Help

	X
1	109
2	130
3	95
4	105
5	120
6	112
7	124
8	108
9	116
10	98
11	122
12	109
13	118
14	111
15	85

Test Variable(s): Acts of Violence [X]

Test Value: 100

FIGURE 12.15 IBM SPSS dialog box when calculating a one-sample *t* after selecting variable and entering hypothesized value.

TABLE 12.4 *IBM SPSS Output from One-Sample* t

```
T-TEST
  /TESTVAL=100
  /MISSING=ANALYSIS
  /VARIABLES=X
  /CRITERIA=CI(.95).
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Acts of Violence	15	110.80	11.833	3.055

One-Sample Test

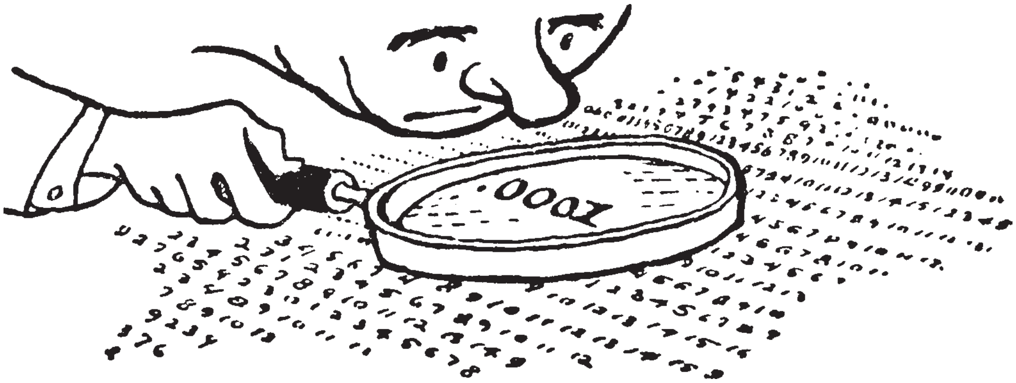
	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	*df*	Sig. (2-tailed)	Mean Difference	Lower	Upper
Acts of Violence	3.535	14	.003	10.800	4.25	17.35

12.16 Levels of Significance versus *p*-Values

Alpha (α) is the probability of rejecting H_0 when H_0 is true. We reject H_0 if $\overline{X}$ falls in the region of rejection based on that level of significance. Although experimenters *should* choose the decision criterion *in advance* of conducting the test (by considering the risk they are willing to take of committing such an error when the sample is drawn), most do not explicitly report the level of significance they chose to use. In fact, some investigators do not always have a precise level of significance in mind prior to examining their results. However, in these cases it can be assumed that the level of significance was no higher than .05, the least stringent criterion allowed by most behavioral science journals.

Although many investigators do not report their selected level of significance, they do report *p*-values. A ***p*-value** is the probability, when H_0 is true, of observing a sample result as deviant or more deviant (in the direction specified in H_A) than the result actually obtained. It is not established in advance and is not a statement of risk; it simply describes the rarity of the sample outcome *if* H_0 is true.

***p*-value**
the probability, when H_0 is true, of observing a sample mean as deviant or more deviant (in the direction specified in H_A) than the obtained value of $\overline{X}$



From *How to Lie with Statistics*, by Darrell Huff, illustrated by Irving Geis. Copyright 1954 and renewed © 1982 by Darrell Huff and Irving Geis. Used by permission of W.W. Norton & Company, Inc.

Look, for example, at Dr. Brown's results for a sample of 100 sixth-grade mathematics achievement scores (Section 12.5). Dr. Brown's obtained sample mean of 90.0 is 2.50 standard errors of the mean greater than the expected value (85.0) if the null hypothesis is true. Only .62% of all the scores in a normal distribution fall at or beyond a score that is 2.50 standard errors above the mean. Thus, for a two-tailed test, the probability of obtaining a sample mean that far beyond the hypothesized mean if H_0 is true is $p = .0062 + .0062 = .0124$. Because this was less than Dr. Brown's selected level of significance, she rejected H_0 and accepted H_A. *If the* p*-value is less than or equal to the level of significance, the sample result is considered sufficiently rare to call for rejecting* H_0.

significant result an obtained sample statistic ($\overline{X}$) that has a low probability of occurring by chance if H_0 is true

Investigators seldom report their sample *p*-values as exact figures. Instead, they commonly report them relative to the landmarks of .05 and .01—and sometimes .001. If an investigator considers his or her result to be a **significant result** (i.e., H_0 is to be rejected), the *p*-value is commonly reported as falling below the landmark, whereas if the result is considered nonsignificant, it will be reported as falling above the landmark. Here are some examples:

	REPORTED *p*-VALUE	
	Investigator considers the results to be	
EXACT *p*-VALUE	*Significant*	*Nonsignificant*
.003	$p < .01$	—
.03	$p < .05$	$p > .01$
.08	—	$p > .05$

In actual research reports, where it is customary to state the obtained statistic (e.g., t_{calc}) and the associated *p*-value, look closely to see whether one-tailed or two-tailed probabilities are intended. Note that in IBM SPSS the *p*-value is reported under the column labeled Sig. (2-tailed) (see Table 12.4). It is the probability of observing a sample mean or one more extreme than it (in either direction in the sampling distributions).

The terminology used by some investigators in describing their results can be quite confusing and can blur the distinction between the *p*-value and the level of significance. An investigator may, for example, report that one set of results was "significant at the .01 level," a second set was "significant at the .001 level," and a third "did not reach significance at the .05 level." Does this mean that $\alpha = .01$, $\alpha = .001$, and $\alpha = .05$, respectively, were used for evaluating the three sets of results? Almost assuredly not. This is just a way of reporting three *p*-values: $p < .01$, $p < .001$, and $p > .05$. Chances are that the level of significance the investigator had in mind, although not explicitly stated, would be the same for evaluating all three sets of results, say, .05.

12.17 Summary

To test a hypothesis about a single population mean using the normal curve model:

Step 1: Define the target population.

Step 2: Specify the null hypothesis, the statistical hypothesis to be tested (H_0). It is always a hypothesis about the population (e.g., H_0: $\mu_X = 100$) and is chosen according to the research question to be studied. The researcher usually hopes to reject H_0.

Step 3: Specify the alternative hypothesis (H_A). We must choose between a directional and a nondirectional alternative, depending on the rationale of the research inquiry.

Step 4: Specify the level of significance (α) to be used as the criterion for decision and determine the z-score value or values that differentiate the region(s) of rejection from the region of retention in the sampling distribution of $\overline{X}$.

Step 5: Decide on n and draw a random sample of that size from the specified population.

Step 6: Calculate $\overline{X}$ and $\sigma_{\overline{X}}$.

Step 7: Calculate in z-score terms the position of $\overline{X}$ in the sampling distribution of the mean that would result by chance when sampling from a population about which H_0 is true:

$$z = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_{\overline{X}}} = \frac{\textit{obtained sample mean} - \textit{hypothesized population mean}\ \textit{(when H}_0\textit{ is true)}}{\textit{standard error of the mean}}$$

Step 8: Reject H_0 if $\overline{X}$ is located in the region of rejection; otherwise, retain H_0.

The difference between $\overline{X}$ and μ_{hyp} depends on random variation and, when the null hypothesis is false, on the additional discrepancy between μ_{hyp} and μ_{true}. Random variation (due to random sampling) is reduced by increasing the sample size.

Although the level of significance should be decided on before any data are collected, many investigators fail to report their α level and instead report only the p-values—the probability of the sample outcome if H_0 is true. In these cases, it is usually safe to assume that the level of significance was no higher than .05.

To test hypotheses about the population mean using z, we must know the population standard deviation, σ_X, to calculate the standard error of the mean, $\sigma_{\overline{X}}$. In actuality, we seldom know the value of σ_X. Without σ_X, we do the next best thing and estimate it from our sample data. However, the sample standard deviation is biased; it tends to underestimate the population parameter. We can obtain a better estimate of σ_X by dividing the sample sum of squares, $\Sigma(X - \overline{X})^2$, by $n - 1$ instead of n. This gives us s_X. When s_X is then divided by $\sqrt{n}$, the result is the estimated standard error of the mean, $s_{\overline{X}}$.

When we substitute $s_{\overline{X}}$ for $\sigma_{\overline{X}}$ in the formula for z, the result is no longer z but a new variable called t:

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}} = \frac{\textit{obtained sample mean} - \textit{hypothesized population mean}\ \textit{(when H}_0\textit{ is true)}}{\textit{estimated standard error of the mean}}$$

Although $\overline{X}$ is normally distributed when certain conditions are met (mainly random sampling), t is not. There is a whole family of t distributions, with the shape of each one depending on the number of degrees of freedom. For hypotheses about single means, $df = n - 1$. The most important difference between z and t is that there is a greater concentration of area in the tails of t distributions. Thus, it takes a larger deviation from expectation ($\overline{X} - \mu_{\text{hyp}}$) to reject H_0 using t. However, as sample size increases, the t distributions more closely resemble the z distribution.

Mathematical Note

Note 12.1 Further Comments on Unbiased Estimates of σ^2 (*Ref.*: Section 12.9)

The variance of a population of scores is given by Formula 5.2*a*

$$\frac{\sum^{N}(X - \mu_X)^2}{N}$$

where N is the number of scores in the population. The best unbiased estimate of this quantity (made from a sample) is:

$$\sigma^2_{X(\text{est})} = \frac{\sum^{n}(X - \mu_X)^2}{n}$$

Note that the deviation of each score in the sample is taken from μ_X. You can demonstrate for yourself that this estimate is unbiased by calculating variance estimates according to this formula for each of the 16 samples shown in Table 12.1. You will find that the mean of the 16 estimates is 5.0, precisely the value of σ^2_X.

When μ_X is unknown, it becomes necessary to take the deviation of each score from its own sample mean: $\sum^{n}(X - \overline{X})^2$. Recall (Section 5.7) that the sum of the squares of deviation scores is a minimum when the deviations are taken from the mean of *that set of scores*, that is, $\overline{X}$. But we should, if we can, compute $\sum^{n}(X - \mu_X)^2$. Because $\overline{X}$ and μ_X will not ordinarily be identical, $\sum^{n}(X - \overline{X})^2 < \sum^{n}(X - \mu_X)^2$, and therefore $S^2_X < \sigma^2_{X(\text{est})}$. To obtain an unbiased estimate of σ^2_X, $\sum^{n}(X - \overline{X})^2$ must be divided by $n - 1$ rather than by n. This adjustment is incorporated into Formula 12.2:

$$s^2_X = \frac{\sum^{n}(X - \overline{X})^2}{n - 1}$$

One interesting point remains. Although s^2 is an unbiased estimator of σ^2, s still tends to underestimate σ (because a square root is not a linear transformation); however, s is an improved estimate. This does not cause important trouble in testing hypotheses about means. First, the degree of underestimation is quite small for samples in which $n > 10$. Second, when the sample size is small, the procedures of this chapter take *exact* account of the substitution of s for σ, so no error is involved.

Key Terms, Concepts, and Symbols

statistical inference (p. 184)
random sampling distribution of $\overline{X}$ (p. 185)
null hypothesis (p. 185)
H_0 (p. 185)
alternative hypothesis (p. 185)
H_A (p. 185)
level of significance (p. 186)
α (p. 186)
region of rejection (p. 187)
region of retention (p. 187)
critical values (p. 188)
rejecting the null hypothesis (p. 189)
retaining the null hypothesis (p. 189)
nondirectional (two-tailed) test (p. 190)
directional (one-tailed) test (p. 191)
observer bias (p. 193)
unbiased estimator (p. 194)
s^2_X, s_X (p. 194)
estimated standard error of the mean (p. 194)
$s_{\overline{X}}$ (p. 194)
t (p. 196)
Student's distribution of t (p. 196)
degrees of freedom (p. 197)
df (p. 197)
statistical conclusion (p. 200)
research conclusion (p. 200)
p-value (p. 206)
significant result (p. 207)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. State whether each of the following could be a null hypothesis: (a) $\mu_X \neq 12$; (b) $\overline{X} > 345$; (c) $\overline{X} = 67$; (d) $\mu_X < 89$; (e) $\overline{X} < 123$; (f) $\mu_X = 45$; (g) $\mu_X > 678$; (h) $\overline{X} \neq 910$.

2. State whether each of the items in Problem 1 could be an alternative hypothesis, and if so, whether it would indicate a one-tailed or a two-tailed test.

3. Why should the choice of a one- or two-tailed test be made before the data are gathered?

4. Why is it inadequate to choose a one-tailed test on the basis that we expect the outcome to be in that direction?

5. In a two-tailed test conducted at the .01 level of significance, we have rejected the null hypothesis. Say whether each of the following is true or false. (a) The probability that the null hypothesis is false is .99. (b) The probability that we made the wrong decision on the null is .01. (c) If we always use the .01 level of significance in our hypothesis testing, over the long run we will make a wrong decision on the null hypothesis 1% of the time. (d) If we always use the .01 level of significance, over the long run we will make the wrong decision on the null hypothesis on 1% of those occasions when we reject the null. (e) If we always use the .01 level of significance, over the long run we will make a wrong decision on the null hypothesis on 1% of those occasions when we retain the null. (f) If we always use the .01 level of significance, over the long run we will make a wrong decision on the null hypothesis on 1% of those occasions when the null is true.

6. Mr. Ratigan wishes to compare the performance of first-grade students in his district with the national norm of 100 on a widely used IQ test. The results for a random sample of his first graders lead him to retain H_0: $\mu = 100$ ($\alpha = .01$) for his population. He concludes, "My research proves that the average first grader in our district falls right on the national norm of 100." Any comment? Compare what we learn when we reject H_0 with what we learn when we retain H_0.

7. An experimenter used $\alpha = .05$ for a statistical test and retained the null hypothesis. Would it be correct to say that the experimenter can be 95% sure that H_0 is true? Explain.

8. The symbolic expression "$p < .001$" occurs in the results section of a research report. Does this indicate that the investigator used the very conservative level of significance $\alpha = .001$ to evaluate the outcome of that particular test? Explain.

9. For the question posed by Dr. Brown, if the population standard deviation on the mathematics achievement test for sixth-grade students is 10, what will Dr. Brown conclude if she uses a two-tailed test at the .05 significance level and obtains a mean of 87.1 for her sample of 100 students?

10. Assume the same data as are given in Problem 9. Suppose that the alternative hypothesis reads H_A: $\mu_X < 85$. Test the null hypothesis at the .05 significance level and state your conclusions.

11. The eighth-grade national norm for a social science test is a score of 123. The research director of a school district wants to know how the scores of pupils in his district compare with this standard. He selects a random sample of 81 students and finds that $\overline{X} = 117$. Assume $\sigma_X = 10$. (a) State the null hypothesis and alternative hypothesis best suited to the nature of his inquiry. Test the null hypothesis (b) at the .05 level of significance and state your conclusions and (c) at the .01 level of significance and state your conclusions.

12. Two years ago, a complete survey of all students in a large university indicated that the mean number of cigarettes smoked per day by male students was 8.3 with a standard deviation of 3.7. The director of student health services wishes to determine whether or not smoking habits among male students at the university have changed over the 2-year period. She obtains the following results for the men from a recently selected random sample of students (in cigarettes smoked per day): $\overline{X} = 7.7$, $n = 108$. (a) Set up H_0 and H_A. (b) Perform the statistical test ($\alpha = .05$). (c) Draw final conclusions.

13. Suppose for Problem 12 that the director's purpose in carrying out the recent study was to check on the effectiveness of an antismoking campaign at the university, instituted shortly after the completed survey of 2 years ago. (a) Set up H_0 and H_A. (b) Perform the statistical test ($\alpha = .05$). (c) Draw final conclusions. (d) Compare your conclusions with those from Problem 12. What does this suggest about which is more likely to give significant results: a two-tailed test or a one-tailed test (provided the direction specified in H_A is correct)?

14. The chancellor at City University is worried that the grades have fallen in the past year. The grade point average for graduating students in the previous 5 years was 2.80 with a standard deviation of 0.25. The chancellor randomly samples 10 seniors from this year's graduating class and obtains the following GPAs: 2.40, 2.80, 2.65, 2.15, 2.65, 2.72, 2.42, 2.70, 2.52, 2.49. What can the chancellor conclude? Use a two-tailed test at the .05 significance level.

15. From Table D in Appendix D, identify (a) the value of t that for 15 degrees of freedom is so high that only 1% of ts would be higher, (b) the centrally located limits for 8 degrees of freedom that would include 95% of ts, (c) for 25 degrees of freedom, the proportion of ts that would be less than $t = -1.316$, (d) for 20 degrees of freedom, the probability in random sampling of obtaining $t \geq +2.6$.

16. We adopt $\alpha = .05$ and test the hypothesis H_0: $\mu_X = 50$. What conclusion should we draw if (a) $n = 10$, $t_{\text{calc}} = +2.10$, and H_A: $\mu_X \neq 50$? (b) $n = 20$, $t_{\text{calc}} = +2.10$, and H_A: $\mu_X \neq 50$? (c) $n = 10$, $t_{\text{calc}} = +2.10$, and H_A: $\mu_X > 50$? Show the critical value of t for each part.

17. Given the scores 2, 4, 4, 7, 8: (a) Calculate s_X by the raw-score formula. (b) Calculate t required to test H_0: $\mu_X = 8.0$ versus H_A: $\mu_X \neq 8.0$. (c) Evaluate t for $\alpha = .05$ and state your conclusions. (d) Evaluate t for $\alpha = .01$ and state your conclusions.

18. The PTA at Central High School is concerned that students who graduate from the school do not score as well on mathematics achievement tests, on average, as do students from other schools in the district. To address their concerns, the school principal randomly selects 15 students from the graduating class and administers the standardized mathematics achievement test used in that school district. The mean score for graduating students in the school district is 98. The scores of the 15 students from Central High are as follows:

105 98 101 110 96 103 104 101
98 105 112 95 105 100 108

(a) State formally the hypotheses necessary to conduct a nondirectional test. (b) Complete the test at the .05 level of significance and state your conclusion.

19. (a) For the data in Problem 18, state H_A if a one-tailed test had been used to address the concerns of the PTA. (b) What would you have concluded with this one-tailed test? Explain.

20. A new brand of automobile tire is advertised to outlast the leading brand, which has a mean lifetime of 80,000 miles. A consumer agency randomly selects six tires from the manufacturer's first production run and tests the tires on a machine that guarantees even tread wear. The results are as follows:

78,000 87,000 85,000 79,000 80,000 86,000

(a) State formally the hypotheses necessary to conduct a nondirectional test. (b) Determine the critical value of t ($\alpha = .05$). (c) Calculate t. (d) State your conclusion. (Do these results cast doubt on the advertised claim?) Hint: How can you simplify the calculations for this problem?

21. Another consumer agency tests (in exactly the same manner) tires from the same batch of tires tested by the agency in Problem 20. It randomly selects 25 tires and obtains $\overline{X} = 82{,}200$ and $s_X = 900$. (a) Calculate t. (b) What does the second consumer agency conclude ($\alpha = .05$)? (c) The mean mileage in the second test was less than in the first. How do you account for the different conclusions?

22. A training director for a large company has been told that, on completion of the training course, the average score of her trainees on the final evaluation should be 100. Her only concern is whether she will have to begin remedial steps to ensure that the population of trainees is not below standard. She draws a random sample of 10 scores of recent trainees: 94, 98, 101, 90, 86, 102, 95, 100, 98, and 92. (a) State the null and alternative hypotheses best suited to the nature of her inquiry. Test the null hypothesis (b) at the .05 level of significance and state your conclusions and (c) at the .01 level of significance and state your conclusions. (d) What is the probability of obtaining through random sampling a sample mean as low or lower if μ_X is as stated in the null hypothesis?

23. Dr. Jones has created a new drug that she believes raises intelligence. To test the drug's effect, she randomly selects 15 rats and tests their learning ability in a maze after they have been injected with the drug. Prior studies indicated that the mean time to complete the maze is 90 seconds. The times to complete the maze for Dr. Jones's subjects are: 95, 84, 80, 92, 90, 75, 100, 88, 94, 82, 78, 94, 86, 85, 83. What can Dr. Jones conclude? Use a one-tailed test at the .05 significance level.

24. The chancellor at City University is worried that the grades have fallen in the past year. The grade point average for graduating students in the previous 5 years was 2.80. The chancellor randomly samples 10 seniors from this year's graduating class and obtains the following GPAs: 2.40, 2.80, 2.65, 2.15, 2.65, 2.72, 2.42, 2.70, 2.52, 2.49. (a) What can the chancellor conclude? Use a two-tailed test at the .05 significance level. Compare this problem with Problem 14. (b) Why was Problem 14 solved with a z test and this problem with a t test?

25. An industrial/organizational psychologist used a new training program designed to improve basic work skills. The training program is typically judged as effective if average scores on a basic work skills test exceed 80. The industrial/organizational psychologist obtained a random sample of participants who recently completed the training program, then administered to them a work skills test. Their scores on the test are given in Data 12A. (a) State the null and alternative hypothesis best suited to the nature of the psychologist's inquiry. (b) Using IBM SPSS, test the null hypothesis at the .01 level of significance and state your conclusions.

DATA 12A

83	85	87	84	94	82
78	81	87	94	97	86
85	85	82	84	84	88
85	88	91	77	78	82
82	84	75	86	90	75
74	86	85	89	82	77
79	87	82	83	86	94
78	87	92	89	80	79
88	82	87	90	88	88
80	87	83	82	89	89

CHAPTER 13

Interpreting the Results of Hypothesis Testing: Effect Size, Type I and Type II Errors, and Power

When you have finished studying this chapter, you should be able to:

- Understand that to say a result is statistically significant does not mean it is important;
- Calculate effect size and understand how it differs from a *p*-value;
- Discuss the two types of errors in hypothesis testing;
- Explain what is meant by the "power of a test";
- List and discuss five factors that influence the power of a test;
- Estimate the sample size necessary to discover the smallest discrepancy (between μ_{hyp} and μ_{true}), if it exists, that we are interested in; and
- Understand the limitations in drawing conclusions about a population if one does not use a random sample drawn from that population.

In Section 12.16 you learned that when the outcome of a statistical test is reported in research literature, it is common to see statements such as "the outcome was significant at the .05 level." What does it mean to say that "the outcome was significant at the .05 level"? In the more exact

significant result
an obtained sample statistic ($\overline{X}$) that has a low probability of occurring by chance if H_0 is true (and thus H_0 is rejected)

language that you have recently learned, this usually means that a null hypothesis and an alternative hypothesis were formulated, the decision criterion was $\alpha = .05$, and the evidence from the sample led to rejection of the null hypothesis.

The use of the word *significant* in connection with statistical outcomes is unfortunate. In common English it implies "important." However, in statistics, a **significant result** means only that an obtained result had a low probability of occurring by chance (that is, the sample value was not within the limits of sampling variation expected under the null hypothesis). Whether the difference between what is hypothesized and what is true is large enough to be important is another matter.

13.1 A Statistically Significant Difference versus a Practically Important Difference

To test the hypothesis H_0: $\mu_X = 100$, we calculate t as

$$t = \frac{\overline{X} - \mu_{\text{hyp}}}{s_X/\sqrt{n}}$$

If t is large enough, the result will be statistically significant and we will reject the null hypothesis. Remember, as applied to the results of a statistical analysis, *significant* is a technical term with a precise meaning: It means that H_0 has been tested and rejected according to a given decision criterion—nothing more, nothing less.

Look again at the formula for t. You can see that the magnitude of t depends both on the quantity in the numerator *and* on the quantity in the denominator. In the numerator, the difference between $\overline{X}$ and μ_{hyp} depends on two things: (1) random variation and (2) when the null hypothesis is false, the difference between μ_{hyp} and μ_{true}. Other things being equal, the larger the difference between μ_{hyp} and μ_{true}, the larger the value of t.

The denominator is a measurement only of random variation (due to random sampling). Ordinarily, if the sample size is very large, the denominator $s_X/\sqrt{n}$ will be quite small. In this event, a relatively small difference between $\overline{X}$ and μ_{hyp} may produce a value of t large enough to lead us to reject the null hypothesis. We may have a result that is "statistically significant," but in which the difference between μ_{true} and μ_{hyp} is so small as to be unimportant. To determine whether your result is important, the simplest remedy is to look at how much difference exists between μ_{hyp} and the obtained sample mean. Is the size of the difference large enough that it matters?

Recall, for example, Dr. Brown's problem (Chapter 12). She asked whether the mean mathematics achievement score of sixth graders in her school district could be 85. To study this question, she drew a sample of 100 students. Suppose someone in the U.S. Department of Education asked the same question about sixth graders in general and drew a random sample of 10,000 students. If we assume that the estimated standard deviation of test scores is the same as the standard deviation in Dr. Brown's case (20), the estimated standard error of the mean is $s_{\overline{X}} = 20.0/\sqrt{10{,}000} = .20$. Adopting $\alpha = .05$ (as before), the Department of Education's researcher will reject the hypothesis that the mean is 85 if his or her calculated t is 1.961 or larger. This value will be reached when the sample mean differs from μ_{hyp} by only about 0.39 points (i.e., an $\overline{X}$ of 85.39 or greater or 84.61 or smaller). Although a value this discrepant indicates that the mean of the population of sixth graders is not 85, a difference of this order is probably not worth bothering about. We say "probably" because *the question of whether the size of the difference is important or not depends on the research question, not the statistical question.* For some research questions, a very small obtained difference between $\overline{X}$ and μ_{hyp} could have important consequences. On the

statistically significant result
does not mean that the result is important

Point of Controversy

The Failure to Publish "Nonsignificant" Results

Because they are limited to a yes or no decision (null hypothesis rejected or null hypothesis retained), some researchers regard results with $p > .05$ as unimportant and those with $p < .05$ as important (and regard highly significant results, i.e., those with very low *p*-values, such as $p < .001$, as very important). Although textbooks strive to convince the novice that significance $\neq$ importance (rejecting the null hypothesis says nothing about the magnitude of the effect—the real difference between μ_{hyp} and μ_{true}), journal editors have translated the yes–no statistical decision into publishable–nonpublishable, thereby bestowing a significant result with an even greater kind of importance. Doctoral degrees are delayed, academic positions and promotions are lost, and grant proposals remain unfunded all because researchers have obtained nonsignificant and, consequently, nonpublishable results.

The consequences to research of journals refusing to publish nonsignificant results are many. For one, this policy encourages the publication of "false positives." For example, if a null hypothesis is true and 20 researchers are independently testing it using $\alpha = .05$, by chance we would expect 1 of the 20 to obtain a significant difference. Though he or she has made an error, this result will probably end up in print while the remaining 19 remain buried in file drawers. The reader of scientific literature will know only of the former. How often this occurs is unknown, but it does happen. Recall from Chapter 1, for example, the studies with cannibal planaria (see Section 1.7).

You might think that the replication of results would serve as a safeguard against incorporating false positives into scientific theory. It would indeed, but journals also tend to emphasize original studies at the expense of replications (Rosnow & Rosenthal, 1989b). In most journals, replications are rarely published (Mahoney, 1976; Sterling, 1959). This bias against replication was echoed by the director of the Federal Trade Commission's Bureau of Consumer Protection in 1982: "If additional substantiation will probably just confirm what the evidence already suggests, then added evidence will not be useful" (*Science,* December 24, 1982, p. 1292).

Perhaps the biggest price humanity pays for this journal policy is that scientific theory is evolving through inductive inference achieved primarily by rejecting null hypotheses at the .05 level (Cohen, 1990). To counter this, the editor of the journal *Applied Psychological Measurement* offered a modest proposal (Weiss, 1989). Rather than submit finished manuscripts to journals, scientists and educators will submit detailed research designs in advance of collecting any data. Journal editors will then accept or reject the design. If they accept it, they will agree to publish the results even if they prove to be statistically nonsignificant. In fact, in 2005 the *New England Journal of Medicine* and 10 other medical journals decided that they would not publish results of clinical studies unless researchers had registered in a publicly available database *before* they began collecting data. This ensures that others would know of the study even if the results are not statistically significant (Dingfelder, 2004).

other hand, if the research question is a trivial one, even a very large obtained difference between $\overline{X}$ and μ_{hyp} is of no practical importance.

What about the statistical test when samples are quite small? In this case, the standard error of the mean will be relatively large, and it will be difficult to discover that the null hypothesis is false, if indeed it is, unless the difference between μ_{true} and μ_{hyp} is quite large. This suggests that an important consideration in experimental design is the size of the sample that should be drawn. You will learn how to determine optimum sample size for tests of hypotheses about means in Section 13.11.

13.2 Effect Size

In 1986, *The New York Times* reported a study that found a "definite" link between children's height (adjusted for age and sex) and IQ. The study was conducted with 14,000 children and the results were "highly significant."

Do these results mean that you should run out and get growth hormone injections for yourself and your children? Before you do, consider the following. If you assume that "highly significant" means $p < .001$, with 14,000 subjects you need a correlation of only .0278 to reach this level of significance. If you further assume a causative relationship, with this low degree of relationship, raising your child's IQ from 100 to 130 would require that you inject enough growth hormone so that his or her height would increase by 14 feet (example and calculations from Cohen, 1990).

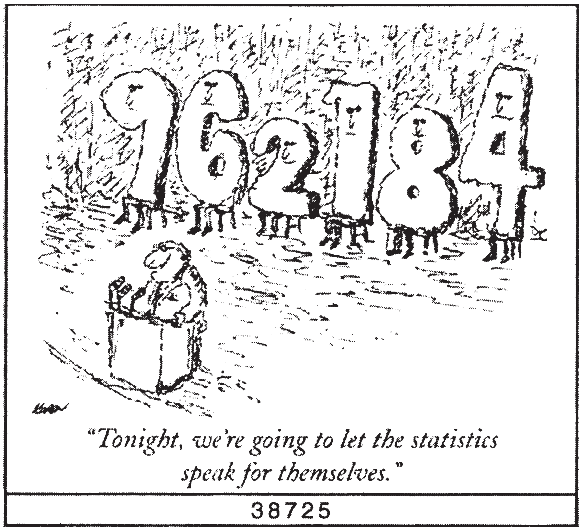
This example demonstrates how foolish it is to rely exclusively on ***p*-values** when evaluating statistical results. With a large enough sample size, we can detect even a very small difference between the value of the population mean stated in the null hypothesis and the true value, but the difference may be of no practical importance. Conversely, with too small a sample size, a researcher may have little chance to detect an important difference.

***p*-value**
the probability, when H_0 is true, of observing a sample mean as deviant or more deviant (in the direction specified in H_A) than the obtained value of $\overline{X}$

How much of a difference between μ_{hyp} and μ_{true} is so great that we want to be reasonably certain of discovering it? To put it another way, some differences between μ_{hyp} and μ_{true} are so small that if the null hypothesis is falsely retained, the error is of no practical concern. For example, suppose a school principal tests the null hypothesis that $\mu_{\text{IQ}} = 100$ for her population of schoolchildren. If she fails to reject the null hypothesis when the mean IQ of her population is really 103, she might feel that the error is of no practical consequence. On the other hand, she would definitely not want to retain that hypothesis if the mean of her population is really 110.

Now, it is fair to say that very likely, all null hypotheses are false. The null hypothesis is stated as a point value, and it is most unlikely that any real population has, for example, a mean IQ of *exactly* 100. *The value of hypothesis testing is that it can be used to reveal differences between* μ_{hyp} and μ_{true} *that are large enough that we care about them.*

The decision as to just how big a difference between μ_X's hypothesized value and its true value must be to be considered important is fundamentally a research question and not a statistical one. It can be made only by a person who knows the meaning of a given difference



"Tonight, we're going to let the statistics speak for themselves."

38725

The Cartoon Bank (2004) Psychology Catalog.

in the context of the question that gave rise to the study. For general use, it is convenient to express the magnitude of the difference between μ_{hyp} and μ_{true} in terms of the number of standard deviations of the measure under consideration (Cohen, 1965, 1988). We refer to such a measure as the **effect size**, symbolized as $\boldsymbol{d}$. For tests of hypotheses about single means,

effect size an estimate of the degree to which the treatment effect is present in the population, expressed as a "number . . . free of the original measurement unit" (Cohen, 1988)

$$d = \frac{\mu_{\text{true}} - \mu_{\text{hyp}}}{\sigma} \qquad (13.1)$$

For practical purposes the formula becomes $(\overline{X} - \mu_{\text{hyp}})/\sigma_X$.

Note that Cohen's d expresses the difference relative to the standard deviation of the set of measures, not the standard error. If we do not know σ, which is almost always the case, we can estimate it by s, giving an estimated d called $\boldsymbol{g}$ (Hedges, 1981; Hedges & Olkin, 1985):

(Cohen's) *d* a measure of effect size; expresses the difference between $\overline{X}$ and μ_{hyp} relative to the standard deviation of the population: $(\overline{X} - \mu_{\text{hyp}})/\sigma_X$

$$\text{estimated } d = g = \frac{\overline{X} - \mu_{\text{hyp}}}{s_X} = \frac{\overline{X} - \mu_{\text{hyp}}}{\sqrt{\dfrac{SS_x}{n-1}}} \qquad (13.2)$$

As an example, look again at Professor Telley's question (Section 12.13). Recall that $\mu_{\text{hyp}} = 100$, $\overline{X}_{\text{obtained}} = 120$, and $s_X = 35.5$. Thus, $g = (120 - 100)/35.5 = +.56$. If $g = +.56$, this means that the difference between the true value and that hypothesized is estimated to be 0.56 of the standard deviation of the variable measured. The positive sign means that the true value is greater than the hypothesized value.

(Hedges') *g* a measure of effect size; expresses the difference between $\overline{X}$ and μ_{hyp} relative to s_X, the estimated population standard deviation: $(\overline{X} - \mu_{\text{hyp}})/s_X$

These two measures (Cohen's d and Hedges' g) are members of what is called the d family of effect sizes. Today, many leading statisticians advocate the use of r for measuring effect size. We will introduce this family of effect sizes in Chapter 14.

Effect size does not tell us the same thing as a p-value does. Highly significant p-values do not indicate that there is a large effect; the study of children's height and IQ with which this section opened is a good example. Conversely, in other studies a nonsignificant result may still be important as judged by the effect size.

> Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1%, we do not . . . ever need to lose sight of the exact strength which the evidence has in fact reached. (Ronald A. Fisher, 1960, p. 25)

Jacob Cohen (1988, 1992) has suggested that *small, medium, and large effects may be defined as corresponding to values of* d *of .2, .5, and .8, respectively.* This view has been adopted by others as a useful convention for evaluating studies. However, it is best considered as a general rule of thumb; *importance can properly be judged only relative to the circumstances of the particular setting.*

effect size (*d* and *g*) useful convention: small = .2 medium = .5 large = .8

Here is an example. In 1988, the Steering Committee of the Physicians' Health Study Research Group reported that one aspirin taken every day reduced the rate of heart attack in one group compared to another group that did not take aspirin. There were approximately 11,000 subjects in each group, and the obtained r of $-.034$ between taking aspirin and heart attacks was highly significant ($p < .00001$). The correlation is quite low and can account for only 0.1% of the variability in the results (Section 9.5). However, the effect size for this type of study (calculated differently from the single-group design in this chapter) was described as "far from unimpressive" (Rosnow & Rosenthal, 1989b). The group that took aspirin had 72 fewer heart attacks and 13 fewer deaths. Generalized to the entire population, that means that the low-cost practice of taking an aspirin daily might result in thousands of fewer heart attacks each year.

In 1999, the American Psychological Association's Task Force on Statistical Inference (Wilkinson et al., 1999) stated that authors should:

> Always present effect sizes for primary outcomes. . . . We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. (p. 599)

The fifth edition of the *Publication Manual of the American Psychological Association* (2001) also emphasizes the need for reporting effect size. The APA already required that authors provide the symbol of the statistic (such as t), the degrees of freedom, the alpha level selected, the obtained value of the statistic, and the p-value. It is probably only a matter of time before most journal editors require that persons submitting manuscripts include some estimated measure of the effect size (some editors already do require it). In the meantime, we urge you to include such measures in your reports. After all, the most important product of a statistical test is not whether one rejects the null hypothesis, which we all know is probably false anyway, but the magnitude of the difference between μ_{hyp} and μ_{true}.

13.3 Errors in Hypothesis Testing

There are, so to speak, two "states of nature": Either the null hypothesis, H_0, is true or it is false. Similarly, there are two possible decisions: we can reject the null hypothesis or we can retain it. Taken in combination, there are four possibilities. They are diagramed in Table 13.1. If the null hypothesis is true and we retain it, or if it is false and we reject it, we have made a correct decision. Otherwise, we have made an error. Notice that there are two kinds of errors. They are called *Type I error* and *Type II error.*

Type I error rejection of a true null hypothesis

If we reject H_0 and in fact it is true, we have made a **Type I error.** Consider the picture of the sampling distribution shown in Figure 13.1 (for simplicity's sake, we consider the sampling distribution of $\overline{X}$ rather than t). It illustrates a situation that might arise if we were testing the hypothesis H_0: $\mu_X = 150$ and using a two-tailed test at the 5% significance level. Suppose our obtained sample mean is 146, which leads us to reject H_0. The logic in rejecting H_0 is that if the null hypothesis is true, a sample mean this deviant would occur less than 5% of the time. Therefore, it seems more reasonable for us to believe that this sample mean came from a population with a mean different from that specified in H_0. By definition, however, the region of rejection identifies a degree of deviance such that even when the null hypothesis is true, 5% of sample means will reach or exceed it. Our sample mean *could* be one of those deviant values obtained through random sampling. Therefore, when

TABLE 13.1 *Two Types of Errors in Hypothesis Testing*

		State of Nature	
		H_0 false	*H_0 true*
Decision	*Retain H_0*	**Type II error**	*Correct decision*
	Reject H_0	*Correct decision*	**Type I error**

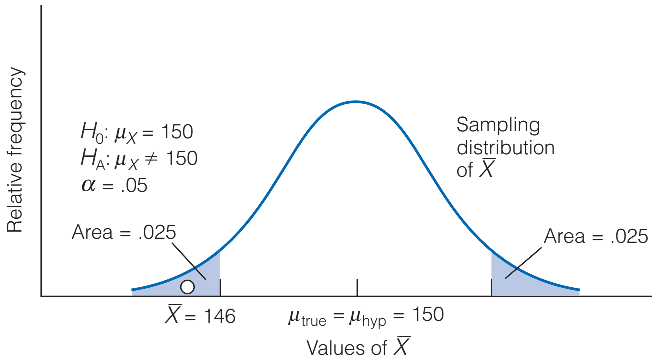

Relative frequency
H_0: $\mu_X = 150$
H_A: $\mu_X \neq 150$
$\alpha = .05$
Sampling distribution of $\overline{X}$
Area = .025
Area = .025
$\overline{X} = 146$
$\mu_{\text{true}} = \mu_{\text{hyp}} = 150$
Values of $\overline{X}$

FIGURE 13.1 Testing H_0: H_0 is true, but $\overline{X}$ leads to a Type I error.

conducting tests using the 5% significance level, and *when the null hypothesis is true*, 5% of sample means will lead us to an incorrect conclusion: rejection of the null hypothesis.

α (alpha) probability of a Type I error

The probability of a Type I error is indicated by the Greek letter **α (alpha)**. More formally:

$$\alpha = \text{Pr (Type I error)} = \text{Pr (rejecting } H_0 \text{ when } H_0 \text{ is true)}$$

Type II error retention of a false null hypothesis

If we retain H_0 and in fact it is false, we have made a **Type II error**. As in the previous example, suppose that we test the null hypothesis that $\mu_X = 150$. Again, we adopt the 5% level of significance. We draw a sample and find its mean to be 152. Now it may be that, unknown to us, the mean of the population is really 154. This situation is pictured in Figure 13.2. In this figure, two sampling distributions are shown. The true distribution is shown with a dashed line and centers on 154, the true population mean. The other is the sampling distribution of $\overline{X}$ that would occur if H_0 is true; it centers on 150, the hypothesized mean. The sample mean actually belongs to the true sampling distribution. But to test the hypothesis that $\mu_X = 150$, we evaluate the mean according to its position in the sampling distribution shown with a solid line. Relative to *that* distribution, it is not deviant enough (from the mean of 150) for us to reject H_0. Our decision will therefore be to retain the null hypothesis that $\mu_X = 150$. It is, of course, an incorrect decision; we have committed a Type II error. To put it another way, we have failed to claim that a real difference existed when in fact it did.

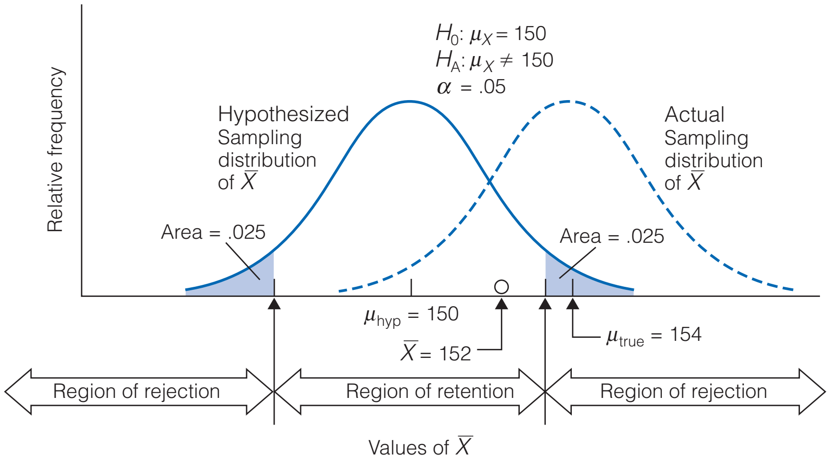

Relative frequency
H_0: $\mu_X = 150$
H_A: $\mu_X \neq 150$
$\alpha = .05$
Hypothesized Sampling distribution of $\overline{X}$
Actual Sampling distribution of $\overline{X}$
Area = .025
Area = .025
$\mu_{\text{hyp}} = 150$
$\overline{X} = 152$
$\mu_{\text{true}} = 154$
Region of rejection
Region of retention
Region of rejection
Values of $\overline{X}$

FIGURE 13.2 Testing H_0: H_0 is false, but $\overline{X}$ leads to a Type II error.

β (beta) probability of a Type II error

The probability of committing a Type II error is indicated by the Greek letter **β (beta)**:

$$\beta = \Pr\ (\text{Type II error}) = \Pr\ (\text{retaining } H_0 \text{ when } H_0 \text{ is false})$$

It is common for researchers to evaluate the outcome of tests according to the .05 or .01 level of significance. However, it is both possible and permissible to use values such as .02 or .001. How should we choose the level of significance? Should α be .05 or should it be .01, or perhaps some other value? Suppose we were to choose $\alpha = .25$. If the null hypothesis is true, now 25% of sample means will fall in the region of rejection we have established, so one test in four will lead us to the incorrect conclusion that the null hypothesis is false (a Type I error). The risk of committing this error seems uncomfortably high.

To reduce the risk, we can set α at a lower level. Suppose we set it very low, say, .001. Now suppose we conduct a test and obtain a sample mean that is so deviant that its probability of occurrence under the null hypothesis is .002. According to the new criterion adopted, we have insufficient evidence to reject the null hypothesis and therefore must retain it. In this case, we run a substantial risk of retaining the null hypothesis when it is false (a Type II error).

How, then, should we choose the level of significance (α)? Think about the practical consequences of rejecting the null hypothesis when it is true. In Dr. Brown's case (Chapter 12), it would result in concluding that the mean mathematics achievement score of sixth graders in her school district is better than the national norm or that it is worse than the norm—when, in fact, it really is the same. If the conclusion is that the mean score is better, Dr. Brown will "point to the score with pride" when the facts don't justify it; if the conclusion is that it is worse, she will begin changing the mathematics program when it is not necessary. What risk should be accepted for the two taken together? In general, one should translate the error into concrete terms and then decide what level of risk is tolerable.

In the same manner, it is useful to translate the abstract conception of a Type II error into practical consequences. Consider Dr. Brown's problem again. A Type II error would occur if she had retained the null hypothesis that the mean of the population of sixth-grade test scores was 85 when in fact they were really above standard or below standard. How important would it be to avoid this error? Once that judgment is made, it can be taken into account in designing the study. You will learn how to do that in Section 13.11.

For general use, $\alpha = .05$ and $\alpha = .01$ make quite good sense. They tend to give reasonable assurance that the null hypothesis will not be rejected unless it really should be. At the same time, they are not so stringent as to raise unnecessarily the likelihood of retaining false hypotheses. *Whatever the level of significance adopted, the decision should be made in advance of collecting the data.*

13.4 The Power of a Test

In Dr. Brown's problem (Chapter 12), she would be making a correct decision (i.e., rejecting a false null hypothesis) if she claimed that the mean mathematics achievement score of sixth graders in her school district was superior to the national norm; and in fact it really *was* superior. Because the probability of retaining a false null hypothesis is β, the probability of correctly rejecting a false null hypothesis is $(1 - \beta)$:

$$(1 - \beta) = \Pr\ (\text{rejecting } H_0 \text{ when } H_0 \text{ is false})$$

power of the test the probability of rejecting a false null hypothesis; **$1 - \beta$**

The value of $(1 - \beta)$ is called the **power of the test**. Among several ways of conducting a test, the most powerful one is that offering the greatest probability of rejecting H_0 when it should be rejected. Because β and the power of a test are complementary, *any condition that decreases β increases the power of the test, and vice versa.* In the next several sections, we will examine the factors affecting the power of a test (and β).

13.5 Factors Affecting Power: Difference between the True Population Mean and the Hypothesized Mean (Size of Effect)

Suppose that we conduct a two-tailed test of the null hypothesis that $\mu = 100$. If the null hypothesis is true, the sampling distribution of $\overline{X}$ will center on μ_{hyp}, as shown by the distribution drawn in dashed outline in Figure 13.3*a*. This is the distribution from which we determine the region of retention, as shown in the figure. Now, if the null hypothesis is false, the true sampling distribution of $\overline{X}$ will center on some other value, μ_{true}, rather than on μ_{hyp}. In reality, we would not know the value of μ_{true}. However, for purposes of discussion, we will assume possible values for it.

Let us assume, for example, that $\mu_{\text{true}} = 103$ and that the true sampling distribution of $\overline{X}$ is that shown in solid outline in Figure 13.3*a*. These are the values of $\overline{X}$ we will actually get in random sampling. What will lead to rejection of the false null hypothesis in this situation? Clearly, getting a sample mean that falls outside the region of retention will lead us to reject the null hypothesis. Because our sample means will actually be drawn from the true distribution, those means in the shaded area (under the solid outline) are the ones that will lead us to reject the null hypothesis. Power, the probability of rejecting a false null hypothesis, is therefore the probability of obtaining such a mean in random sampling: judging by eye, it appears that this probability is about .2.

Now let us assume that the true population mean and the hypothesized mean are substantially different, say, $\mu_{\text{true}} = 112$. This is illustrated in Figure 13.3*b*. Here, there is little overlap of the two sampling distributions of $\overline{X}$, and most samples have means that will lead to the rejection of H_0. *In general, the greater the difference between* μ_{true} and μ_{hyp}, *the more powerful is the test.*

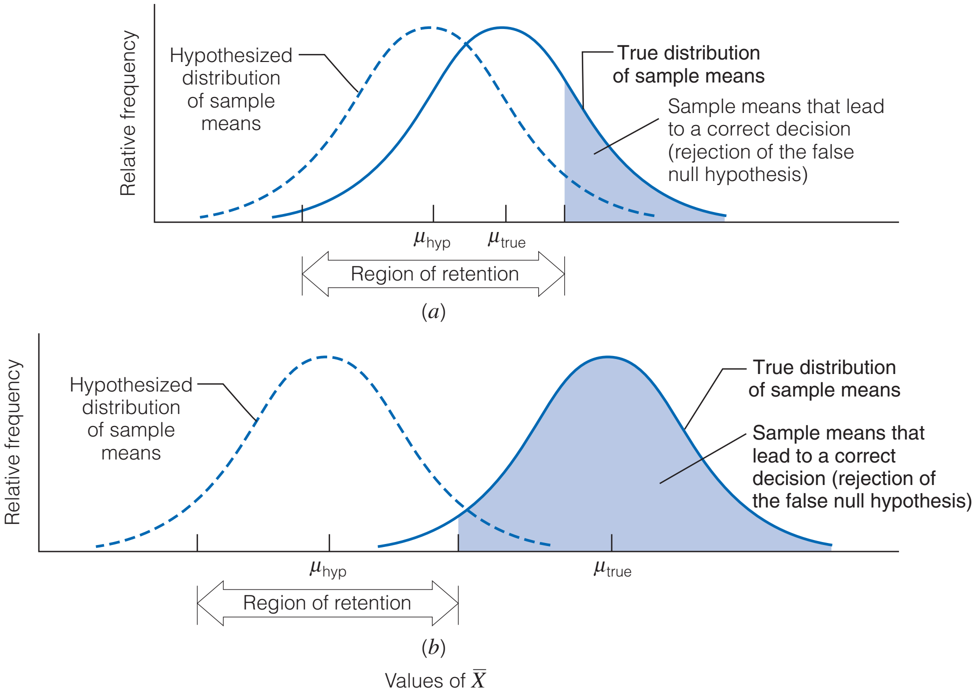

Relative frequency
Hypothesized distribution of sample means
True distribution of sample means
Sample means that lead to a correct decision (rejection of the false null hypothesis)
μ_{hyp}
μ_{true}
Region of retention
(a)
Relative frequency
Hypothesized distribution of sample means
True distribution of sample means
Sample means that lead to a correct decision (rejection of the false null hypothesis)
μ_{hyp}
μ_{true}
Region of retention
(b)
Values of $\overline{X}$

FIGURE 13.3 Probability of rejecting a false null hypothesis as a function of the size of the difference between μ_{hyp} and μ_{true}.

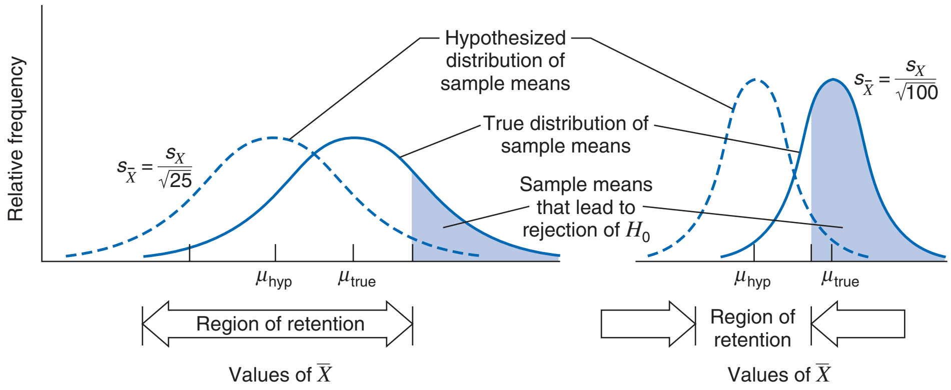


FIGURE 13.4 Probability of rejecting a false null hypothesis as a function of sample size.

13.6 Factors Affecting Power: Sample Size

The probability of rejecting a false null hypothesis is also related to sample size. Suppose, for example, that we consider a sample size of 25. The estimated standard error of the mean is $s_{\overline{X}} = s_X/\sqrt{25} = .2s_X$. If, instead, the sample size were 100, the estimated standard error of the mean would be $s_{\overline{X}} = s_X/\sqrt{100} = .1s_X$, which is only half as large (assuming that the two values of s_X are about equal). In general, the larger the size of the samples, the smaller the standard deviation ($s_{\overline{X}}$) of the sampling distribution of $\overline{X}$. Figure 13.4 shows how the sampling distribution might appear when all factors are the same except for sample size.

When sample size is larger, there is less overlap between the true and the hypothetical sampling distributions of $\overline{X}$. Consequently, the probability is greater of drawing a sample that leads to rejection of the false null hypothesis. *Other things being equal, the larger the sample size, the smaller the standard error of the mean and the greater the power of the test.* The selection of sample size is the simplest method of increasing power (see Section 13.11).

13.7 Factors Affecting Power: Variability of the Measure

In the previous section, we showed that an increase in sample size increases power by reducing the standard error of the mean. Because the estimated standard error of the mean is $s/\sqrt{n}$, *another way to increase power is to reduce the size of* s. At first glance, one might think that this is beyond the control of the investigator. Actually, this opportunity is often open, but advance planning is necessary.

The standard deviation reflects variation attributable to the factors we wish to study, but it also reflects variation attributable to extraneous and irrelevant sources. *Any source of extraneous variation tends to increase* s *over what it would be otherwise*, so an effective effort to eliminate such sources of variation will tend to decrease s and thus increase power.

13.8 Factors Affecting Power: Level of Significance (α)

Power is also related to the choice of α. Figure 13.5*a* shows the probability of rejecting H_0 when it is false and $\alpha = .05$. Figure 13.5*b* shows the identical situation except that $\alpha = .01$. In both diagrams, the sample means represented by the shaded areas in the distributions shown

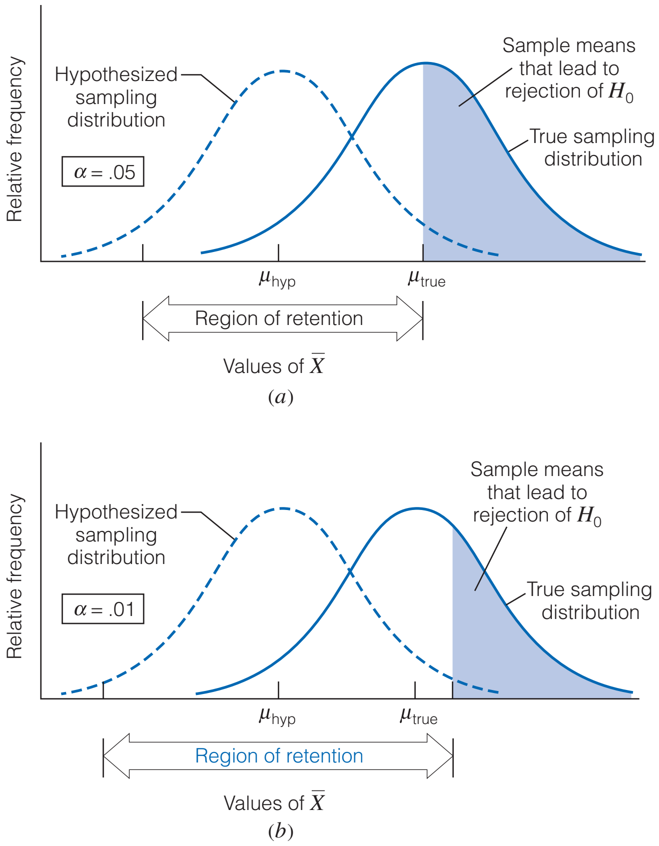


FIGURE 13.5 Probability of rejecting a false null hypothesis as a function of α.

factors affecting power (1) difference between μ_{hyp} and μ_{true}; (2) n; (3) s; (4) level of significance; and (5) choice of a one-tailed or two-tailed test

in solid outline are those that fall in the region of rejection. The proportionate frequency of such means is greater when $\alpha = .05$ than when $\alpha = .01$. Therefore, there is a greater chance of obtaining a sample mean that will lead to the rejection of a false H_0 when α is larger. *In general, reducing the risk of Type I error increases the risk of committing a Type II error and thus reduces the power of the test.*

13.9 Factors Affecting Power: One-Tailed versus Two-Tailed Tests

The power of a test is also affected by the choice of the alternative hypothesis. If we conduct a one-tailed test of the hypothesis H_0: $\mu_{\text{hyp}} = 100$ against the alternative H_A: $\mu_{\text{hyp}} > 100$, we consider H_0 to be false if and only if $\mu_{\text{true}} > \mu_{\text{hyp}}$. Figure 13.6 shows the test of H_0 when α is .05 and $\mu_{\text{true}} > \mu_{\text{hyp}}$. The two diagrams represent identical test conditions except for the alternative hypothesis. In a z test, for example, $z_{\text{crit}} = \pm 1.96$ when H_A is nondirectional and $+1.645$ when H_A is directional. Compare the proportion of the sample means leading to rejection of H_0 (the shaded area) in the two diagrams. It is greater for the one-tailed test. *Other things being equal, the probability of rejecting a false null hypothesis is greater for a one-tailed test than for a two-tailed test (but only when the direction specified by* H_A *is correct).*

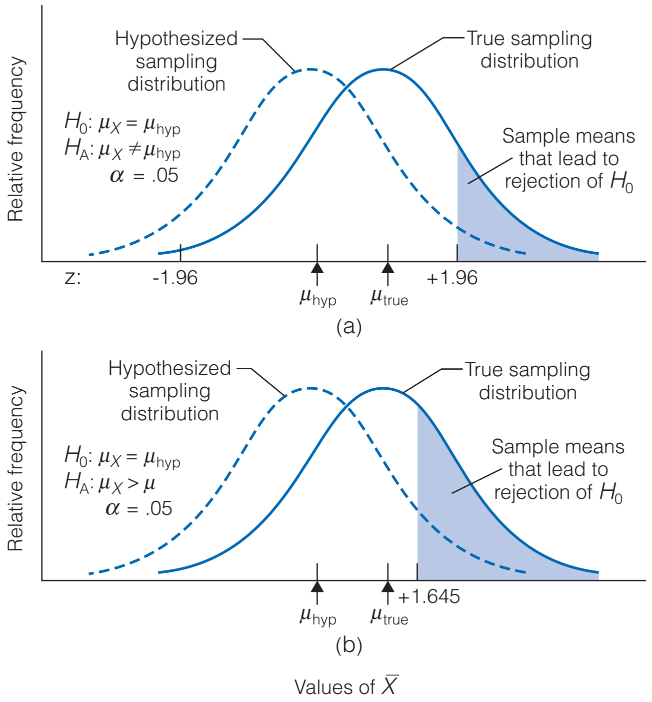


FIGURE 13.6 Probability of rejecting a false null hypothesis in (*a*) a two-tailed test and (*b*) a one-tailed test when $\alpha = .05$.

13.10 Calculating the Power of a Test

In this section you will learn how to calculate power. *Power can be calculated only when the true value of μ can be specified*. Although we do not know this value, we can calculate power for different *possible* values of μ_{true}. The result is a table or graph that allows us to see the consequences of testing hypotheses under various circumstances and from which we can determine the sample size required to keep power at the level we desire.

Suppose that we wish to test the following hypothesis:

$$H_0\text{: } \mu_X = 100$$

$$H_A\text{: } \mu_X > 100$$

$$\alpha = .05$$

For the sake of simplicity, we will work with z rather than t. If the sample size is 25 and we know that $\sigma_X = 20$, then $\sigma_{\overline{X}} = \sigma_X/\sqrt{n} = 4.00$. To test the null hypothesis, we construct the sampling distribution of $\overline{X}$ that occurs when the hypothesis is true. We show this sampling distribution in dashed outline in Figure 13.7. If, relative to this distribution, we obtain a sample mean that falls above the limit $z = +1.645$, we reject H_0.

Suppose that the null hypothesis is false because $\mu_{\text{true}} = 104$. In random sampling, the values of $\overline{X}$ actually obtained will follow the true sampling distribution, shown in solid outline in Figure 13.7. If our obtained sample mean is one of those shown in the shaded portion of this distribution, the decision will be to reject H_0, and we will have made a correct decision. What is the probability of making this correct decision? It is the probability of getting one of the means that falls in the shaded region of the true sampling distribution.

To find the area in the shaded portion of the curve, we must express its boundary in z-score terms. We already know what the boundary is in z-score terms *relative to the hypothesized*

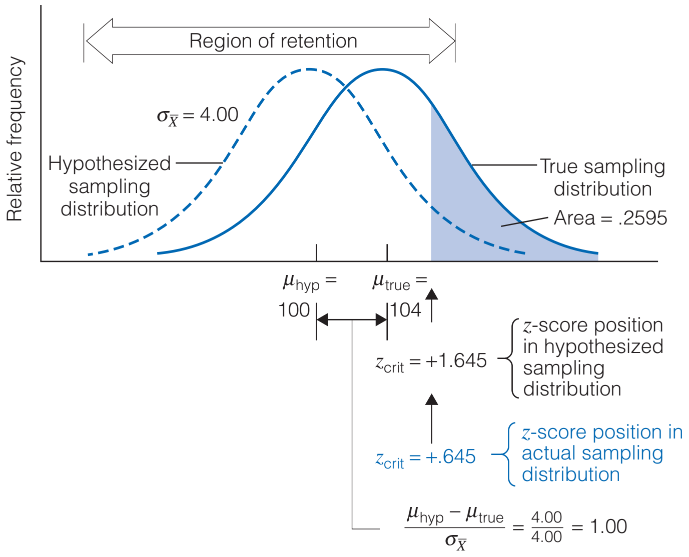


FIGURE 13.7 Probability of rejecting a false null hypothesis when testing the hypothesis that $\mu_X = 100$.

population mean: $z = +1.645$. However, we must know what this z-score position is in the true distribution. Notice the magnitude, taken without regard for direction, of the difference between μ_{hyp} and μ_{true}; it is $|\mu_{\text{hyp}} - \mu_{\text{true}}| = 4.00$. We convert it to z-score units by dividing by $\sigma_{\overline{X}}$:

$$\frac{|\mu_{\text{hyp}} - \mu_{\text{true}}|}{\sigma_{\overline{X}}} = \frac{4.00}{4.00} = 1.00$$

We see that the boundary, which was 1.645 z-score units above μ_{hyp}, is $(1.645 - 1.00)$ units above μ_{true}. *Relative to the true distribution*, then, $z_{\text{crit}} = +.645$. We can now find the area above this boundary using the methods discussed in Section 6.4, assuming the normality of the

Point of Controversy

Meta-Analysis

In this book, we focus on how to summarize and analyze a sample of scores obtained in a single study. It is rare, however, that a single study ever completely answers a research question. For any given topic, there are typically a large number (sometimes hundreds) of studies. Although each may address the same general question, they often differ with regard to type of subjects, size of samples, variables, procedures, and definitions. Eventually, the results of these separate studies must be compared and integrated, and conclusions must be drawn to advance knowledge in a particular area.

The traditional approach has been the "authoritative" review article, in which an "expert" (i.e., someone who has published several papers in the particular area) forms a judgment based on a selection of the available data. But how do these experts weigh the results of different studies? The assessments are qualitative, not quantitative, and often very subjective (generally giving more weight to the reviewer's own research findings and others with similar results). According to one analysis of review articles, experts' opinions are "always dependent on how they are trained, not on the body of evidence" (Chalmers, cited in Mann, 1990). As a result, it is not uncommon for two or more reviews of the same research question to differ in their conclusions.

Attempts to find a quantitative means to compare the results of different studies can be traced back to Ronald A. Fisher in the 1920s. One of the biggest problems is dichotomous hypothesis testing with its emphasis on *p*-values. What does it mean when one study reports results with $p < .001$, another (investigating the same research question with slightly different methodology) reports $p < .05$, and still a third (with slightly different methodology again) reports a failure to reach significance? Even when two studies employ the same methodology, the *p*-values may differ substantially if sample sizes are not the same. In other words, when significance tests are the only tool that researchers use, the results can obscure underlying regularities (Schmidt, 1992). Sections 13.1 and 13.2 pointed out the weakness of comparing *p*-values from different studies and also showed the advantage of using effect size for such comparisons. In fact, effect size measures, particularly the standardized mean difference *d*, have led to a new method of comparing results from different studies. The new methodology, named **meta-analysis** by Glass in 1976, uses formal statistical procedures to combine results from many similar studies.

meta-analysis a quantitative procedure that allows us to combine the numerical results from multiple studies

The basic logic of meta-analysis is easy to understand. *The observational unit is a study*. For each study, the technique compares what was found with the "null" hypothesis—the hypothesis that there is no effect. The effect sizes in the many studies will differ only randomly from zero effect if the null hypothesis is true. In other words, the final result should be near zero if we add all the effect sizes together. However, if there is an effect, the results of most of the studies should be in the same qualitative direction (regardless of differences in subjects, procedures, etc.), and the cumulative result should differ sharply from the null hypothesis (see Garfield, 1991, or Mann, 1990). The goal of meta-analysis is to provide an estimate of the level and variability of the effect size for a particular area of research.

As an example, consider the question of whether a full moon affects behavior. Does a full moon result in an increase in mental and emotional disorders, suicides, crimes, and so forth? Many people apparently believe so, for there were over 40 studies conducted on the subject between 1961 and 1984. Rotton and Kelly (1985) combined these 40+ studies in a meta-analysis. Although some studies had found significant results, when they were quantitatively combined, the effect size was found to be negligibly small.

Meta-analysis is *not* done by reading abstracts or the conclusion sections of research papers, but by extracting and cumulating data. Therein lies one of the biggest problems; effect size is rarely reported in research articles (see Section 13.2) and must be estimated, often from inadequately reported significance levels. Meta-analysts are among the leaders pushing for stricter standards for publication (see Wachter, 1988). Further problems arise from the fact that journals tend to publish only significant results and few replications (see the first Point of Controversy in this chapter), raising the possibility that the procedure may summarize an unrepresentative batch of studies (Hedges & Vevea, 1996; Levine, 2009; Moher & Pham, 1999). Ideally, a summarizing process should include the results of all the studies in a particular area, including those for which $p > .05$. One of the biggest criticisms of meta-analysis is that in dealing with bulk, particularly in an area in which there are a large number of poorly designed or nonexperimental (observational) studies, bad science wins by sheer weight of numbers ("garbage in and garbage out") (Moher & Pham, 1999; Shapiro, 1994).

We must not lose sight of the fact that all of these criticisms of meta-analysis are equally true of the traditional, nonquantitative, "expert" review (Rosenthal & DiMatteo, 2001). With meta-analysis, poor studies are more easily exposed *by comparing effect size* than by informally reading the authors' own summaries of their work. Meta-analysis is not vote counting, and proponents say that good studies should stand up against a larger number of bad studies.

Like any other technique, meta-analysis should be done properly and with caution. "If the meta-analysis is based on randomized experiments, strong causal inferences are often warranted. If the meta-analysis is based on observational studies, causal inferences are as risky as they are in the case of individual observational studies" (Rosenthal & DiMatteo, 2001). Despite some misuses, meta-analysis has already helped focus attention on effect size and the advancement of science as a cumulative process (instead of as a function of reaching some magical *p*-value). We must not forget "that with further trial it (a hypothesis) might come to be stronger or weaker" (Ronald A. Fisher, 1960, p. 25).

sampling distribution. The area between z_{crit} and μ_{true} is approximately .2405 according to the table of areas under the normal curve. Thus, the probability of obtaining a sample mean that falls outside the boundary is $.5000 - .2405 = .2595$, the power of this test. Although the null hypothesis is false (remember, $\mu_{\text{true}} = 104$), our chances of rejecting it are not very good under these circumstances.

In the calculations just shown, we assumed that we knew σ_X. When we estimate it from the sample, the normal curve is only an approximate model, with the approximation growing less accurate as sample size is reduced. Other models are available for these circumstances, but space limitations do not allow us to consider them here. Also, keep in mind that in our example, we used a one-tailed test (because it is easier to show graphically), but computing power in a two-tailed test requires that we consider both tails of the distribution.

13.11 Estimating Power and Sample Size for Tests of Hypotheses about Means

> [Authors of papers submitted for publication should] provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. (Wilkinson & the Task Force on Statistical Inference, 1999)

One of the greatest problems in the scientific literature today is that many studies are conducted with low power. Surveys of some leading psychology journals, for example, have found that the median power to detect a medium effect size is only about .45 or less (Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989). The same is true of studies published in medical journals (Moher, Dulberg, & Wells, 1994). Jacob Cohen, one of the leading advocates of power analysis (1988), recommends that *the desirable level of power be .8* (1965, 1992). With the low levels of power actually used in most studies, researchers often fail to reject the null hypothesis in situations in which H_0 is false and there is a medium or even a large effect size (Hunter, 1997).

desirable level of power recommended to be .8

The easiest way to increase power is to increase sample size. We can determine the amount of power for a particular sample size or, conversely, the sample size required to achieve a desired level of power if we can specify the following:

1. The risk (α level) we are willing to take of rejecting a true null hypothesis.
2. The magnitude of the *smallest* discrepancy that, if it exists, we wish to be reasonably sure of discovering.

Any one of the four variables we have discussed is a function of the other three.

power curves "used to plan studies by ensuring that the power of statistical tests used will be adequate for the smallest effect size deemed to be of practical significance in a given context" (Hedges & Pigott, 2001)

To show the relationship among these variables, let us adopt $\alpha = .05$ and test H_0: $\mu_X =$ some specified value using a two-tailed test. The relationships are shown in Figure 13.8. Effect size is on the horizontal axis and power is on the vertical axis. The **power curves** (as Cohen calls them) show the relationship between effect size and power for a variety of sample sizes. Note that if the power curves were extended to the left to the vertical axis, where $d = 0$, the curves would converge and intersect the vertical axis at .05 because α in this case is set at .05. Intuitively, this makes sense. What should power be when the null hypothesis is true, that is, $d = 0$? It would be α.

Recall Dr. Telley's question (Section 12.13). We estimated d to be $+.56$ (Section 13.2). With 25 subjects, what was the power of her test? To determine this, we locate a point between .5 and .6 on the horizontal axis (effect size) and move vertically as shown by the dashed line to the curve for $n = 25$ and then horizontally to the vertical axis (power), where we find that power is about .78. This means that if the difference between μ_{hyp} and μ_{true} really is $d = +.56$, Professor Telley could expect to reject the null hypothesis and conclude that there is a

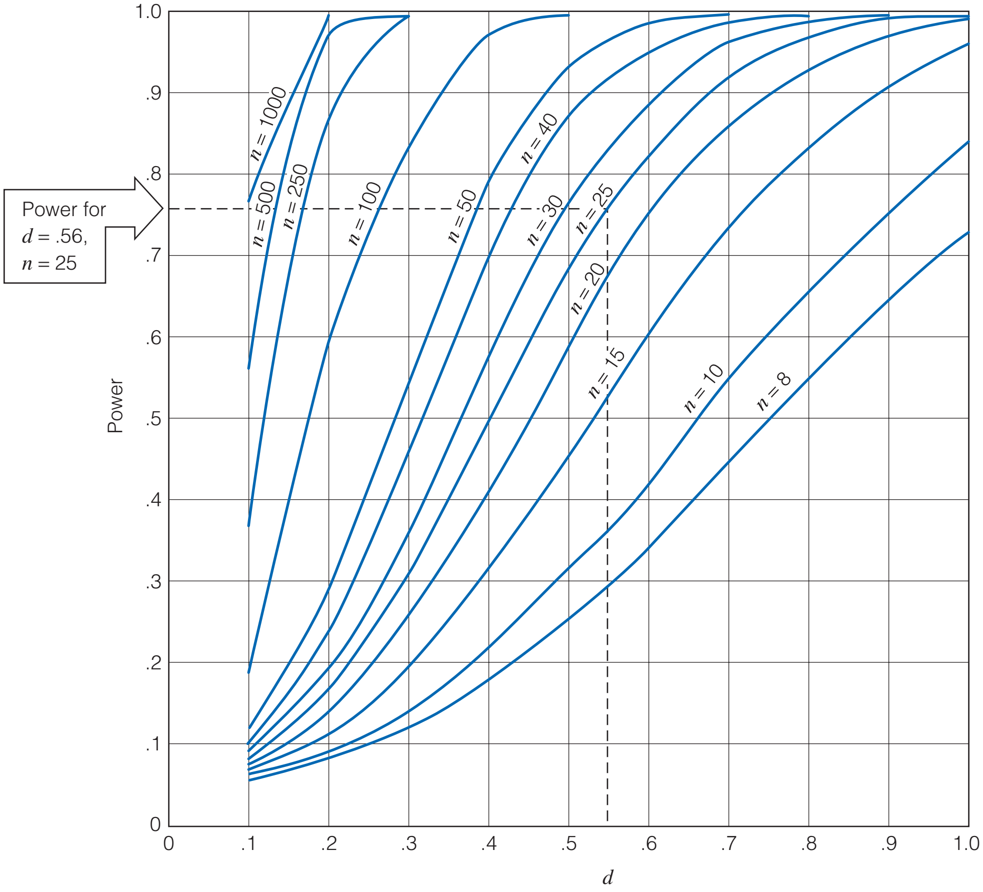


FIGURE 13.8 Power curves for the t test of H_0: μ_X = some specified value; $\alpha = .05$ against a two-tailed alternative. (*Source:* Data for this figure were obtained from J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed., Lawrence Erlbaum Associates, New York, 1988. Used with permission.)

difference in only 78 out of 100 identical replications of her study; in the remaining 22 replications, chance would lead her to retain H_0. Notice that for a smaller effect size, 25 subjects provide even less certainty of detecting a difference.

All too frequently, studies are performed by setting the level of significance and then deciding intuitively (or out of convenience) on sample size. Power curve charts (such as Figure 13.8) make a more rational approach possible. Using the charts, one decides on a desired level of power and then selects sample size *according to the minimum effect size one wishes to discover* (if it exists). For example, if Professor Telley had wanted a probability of .9 of finding a medium-sized effect ($d = .5$) to be significant, a sample size of about 44 would have been required.

Charts like Figure 13.8 are also useful for evaluating the adequacy of published studies. We obtain from the study information about α and the sample size and then select whatever value of power seems appropriate. We may then use the power curve for the particular values of α and n to learn what value of d corresponds to these parameters. (Conversely, we may select an effect size and determine with what level of power the test was conducted.) This will tell us what effect size had a good chance of being discovered if it had existed. Suppose, for example, that a two-tailed test had been conducted with $\alpha = .05$ and $n = 20$. If we assume power = .8, we find .8 on the vertical axis, move horizontally to the curve for $n = 20$, and then move vertically to the horizontal axis to find the effect size. We find that d is about .65. Under the conditions of the test, then, a difference between μ_{hyp} and μ_{true} of about .65 of a standard deviation had a probability of .80 of being detected.

Cohen (1988) provides extensive tables for obtaining precise power values under a variety of circumstances—different α levels and one- and two-tailed tests—as well as other statistical indices (such as correlation coefficients and proportions).

13.12 Problems in Selecting a Random Sample and in Drawing Conclusions

generalizing results of a statistical test we may properly make conclusions only about the specific population from which we drew our sample

The ideal way to conduct statistical inference is to define carefully the target population, identify each element of the population and assign it an identification number, and draw a random sample using a computer or a table of random numbers. In behavioral science, only occasionally do real problems permit these specifications to be met. For example, the theoretical scientist who is studying the learning behavior of "laboratory rats in general" cannot draw a truly random sample and must limit his or her observations to the learning behavior of rats in the latest shipment from Superba Breeding Laboratories. One problem that arises is that there are important differences among strains of laboratory rats. Similarly, if a scientist about to study a problem in learning, motivation, attitude change, perception, or whatever, attempts to define the population about which she would really like to draw conclusions, she will usually be overwhelmed with the difficulties that would lie ahead. The learning scores of all elementary schoolchildren in the United States? All women? All members of a minority group? She pushes these thoughts to the back of her mind and reaches for a convenient group of subjects to begin her exploration. This often turns out to be students currently enrolled in a general psychology class. She does not really wish to draw conclusions about the learning scores of such students, but on statistical grounds, it is proper to generalize the inferential outcome of a test *only* to the population from which the observations may be considered to be a random sample. If wider generalization is considered, it is incumbent on the researcher to repeat the study in the wider population or to make a reasonable showing that there is no relevant difference between the population sampled and the wider population about which it is desired to draw conclusions. The latter is a very difficult thing to do.

statistical conclusion a conclusion about the numerical property of the data (reject or retain H_0)

research conclusion a conclusion about the subject matter

In Section 1.3, the distinction was made between a statistical conclusion and a research conclusion (one about the subject matter). A **statistical conclusion** says something about a parameter of the population, such as μ. But a **research conclusion** says something about the meaning of the study for psychology, or education, or some other discipline. The former can be done as an "automatic" process once statistics is learned. However, moving from the research question to an appropriate statistical question, and finally from a statistical conclusion to a research conclusion, requires the highest qualities of knowledge and judgment on the part of the investigator.

13.13 Summary

The difference between $\overline{X}$ and μ_{hyp} depends on random variation and, when the null hypothesis is false, on the additional difference between μ_{hyp} and μ_{true}. With a large enough sample, it is possible to detect even a very small difference between μ_{hyp} and μ_{true}. In statistics, calling a result "significant" means only that it led the researcher to reject the null hypothesis because it had a low probability of occurring by chance. However, statistical procedures cannot tell us whether the result is important.

The dichotomous nature of hypothesis testing has resulted in undue importance being given to *p*-values. Unfortunately, *p*-values tell us nothing about the effect size. This chapter emphasized the importance of determining *how big* rather than how significant an observed difference is. This is measured by finding effect size, *d* or *g*. Many researchers consider small, medium, and large effects to correspond to values of *d* of .2, .5, and .8, respectively.

At the beginning of a study, *before* any data are collected, the researcher must decide on the level of significance to be used. Setting the level of significance high increases α, the probability of rejecting a true null hypothesis (a Type I error). Setting the level of significance low increases β, the probability of retaining a false null hypothesis (a Type II error).

The power of a test is $(1 - \beta)$, where β is the probability of retaining H_0 when H_0 is false (a Type II error). Thus, β and power are complementary, so any condition that decreases β increases the power of the test, and vice versa.

Power analysis necessitates that researchers consider the size of effect they consider to be important. Among several ways of conducting a statistical test, the most powerful one is that offering the greatest probability of rejecting H_0 when H_0 is false. The probability of committing a Type I error (α) is set by the investigator in advance of collecting the data, but power is affected by a number of factors, including: (1) the magnitude of the difference between what is hypothesized and what is true; (2) the sample size (the larger the sample size, the greater the power); (3) the standard deviation of the variable (the smaller σ is, the greater the power is); (4) the choice of level of significance (reducing α increases β and reduces power); and (5) the choice of a one-tailed or a two-tailed test (power is greater for a one-tailed test if the direction of H_A is correct).

Effect size, sample size, α level, and power are all related; any one is a function of the other three. Thus, if investigators specify the magnitude of the smallest difference that is important for them to discover (if it exists), then they can calculate (based on the selected level of α) the size of the sample(s) necessary to ensure a given level of power for discovering a difference of that size. From knowledge of sample size and significance level, one can also evaluate published studies in terms of their power to discover a particular effect size. Unfortunately, most published studies are conducted with low power.

When moving from the statistical conclusion to the research conclusion, on statistical (mathematical) grounds we can really make conclusions only about the specific population from which we drew our random sample. If wider generalization is desired, we must repeat the study in the wider population or reasonably show that there is no relevant difference between the population sampled and the wider population about which we wish to draw conclusions.

Key Terms, Concepts, and Symbols

significant result (p. 214)
practically important difference (p. 214)
p-value (p. 216)
effect size (p. 217)
d (p. 217)
g (p. 217)
Type I error (p. 218)
α (p. 219)
Type II error (p. 219)
β (p. 220)
power of the test (p. 220)
$1 - \beta$ (p. 220)
meta-analysis (p. 226)
power curves (p. 227)
generalizing results (p. 229)
statistical conclusion (p. 229)
research conclusion (p. 229)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In a 1977 study (Smith & Glass), a correlation of .32 was found between whether or not an individual had psychotherapy and whether or not his or her condition improved. For this study, $r^2 = .10$, meaning that psychotherapy could account for only 10% of the variability in results. Are you impressed with the results? Why or why not?

2. A researcher drew a sample of 400 people, measured their IQs, and tested the null hypothesis that $\mu_{\text{IQ}} = 100$. The researcher obtained $z = +3.00$, which led her to reject the hypothesis (σ was 15.0). (a) What was the standard error of the mean? (b) What was the mean of the sample? (c) The test yielded a statistically significant outcome. Does it appear to be an important one? Explain. (d) How much of a standard deviation did the discrepancy between $\overline{X}$ and μ_{hyp} amount to?

3. Calculate the effect size for the example in Section 13.10.

4. Calculate g for Problem 18 in Chapter 12.

5. Calculate g for (a) Problem 20 in Chapter 12 and (b) Problem 21 in Chapter 12.

6. Calculate g for Problem 22 in Chapter 12.

DATA 13A

H_0: $\mu_X = 80.0$ $\quad$ $\sigma_X = 20.0$
H_A: $\mu_X \neq 80.0$ $\quad$ $n = 100$

7. With regard to Data 13A, (a) if $\alpha = .05$ and $\mu_{\text{true}} = 81.0$, what is β? What is the power of the test? (b) if $\alpha = .05$ and $\mu_{\text{true}} = 82.0$, what is β? What is the power of the test?

8. With regard to Data 13A, (a) if $\alpha = .05$ and $\mu_{\text{true}} = 83.0$, what is β? What is the power of the test? (b) If $\alpha = .01$ and $\mu_{\text{true}} = 83.0$, what is β? What is the power of the test?

9. With regard to Data 13A, (a) if $\alpha = .05$ and $\mu_{\text{true}} = 82.0$, what is β? What is the power of the test? (b) If $\alpha = .05$ and $\mu_{\text{true}} = 82.0$, but $\sigma_X = 10.0$, what is β? What is the power of the test?

10. With regard to Data 13A, (a) if $\alpha = .05$ and $\mu_{\text{true}} = 84.0$, what is β? What is the power of the test? (b) If $\alpha = .05$ and $\mu_{\text{true}} = 84.0$, but H_A: $\mu_X > 80.0$, what is β? What is the power of the test?

11. A psychologist wonders whether performance by students in her class has fallen this semester. Among students in past semesters, mean GPA = 3.1 and $\sigma_{\text{GPA}} = .30$. She decides to adopt $\alpha = .05$ and power = .8. If the difference between present students and those in past semesters is as great as .09 grade point, she would want to know it. (a) State the null hypothesis and alternative hypothesis for a two-tailed test. (b) What should the size of the sample be to test this null hypothesis? Use the power curve chart.

12. In reference to Problem 11, what should sample size be if the psychologist feels that she should choose power = .90 and it is important to know whether the difference is .15 grade point (or more) and power = .95?

13. If you conduct the same experiment 100 times under identical circumstances and the power of your test is .7, how many times would you expect to obtain significant results?

14. Given that $\mu_X = 97$ and $\sigma_X = 4$, if you conduct a two-tailed test of H_0: $\mu_X = 100$ with $\alpha = .05$, how many times in 50 replications would you expect to obtain significant results for samples of size (a) 10? (b) 20? (c) 50?

15. For a two-tailed test of H_0: $\mu_X = 100$ with $\alpha = .05$ and $n = 10$, what is the probability of detecting a d of (a) .2? (b) .5? (c) .8? (d) What does this tell you about using samples of this size?

16. For a two-tailed test of H_0: $\mu_X = 50$ with $\alpha = .05$ and $n = 100$, what is the probability of detecting a d of (a) .2? (b) .5? (c) .8? (d) What does this tell you about using samples of this size (compare with Problem 15)?

17. To see whether Public Health Service attempts to educate the public about the dangers of having multiple sexual partners have been effective, a survey was conducted of the sexual behaviors of women at Brown University in Providence, RI (DeBuono et al., 1990). The researchers found that about the same number of college women were having multiple partners as was found in a 1975 survey. The researchers concluded, "Public health campaigns have not had a substantial influence on the habits and behavior of these well-educated young adults." *The Washington Post* reported, "A study published today found that college women are just as sexually active with multiple partners as they were in the mid-1970s." Comment on the similarities and differences of these two conclusions.

18. Let us now return to the question of statistical significance versus importance. In 1977, studies found that laboratory rats fed huge doses of the noncaloric sweetener saccharin developed bladder tumors. The results were statistically highly significant. By itself, is this strong enough evidence to remove saccharin from the market?

CHAPTER 14

Testing Hypotheses about the Difference between Two Independent Groups

When you have finished studying this chapter, you should be able to:

- Understand the nature of the random sampling distribution of $(\overline{X} - \overline{Y})$;
- Calculate $s_{\overline{X}-\overline{Y}}$ (the estimated standard error of the difference between two means) and t for testing hypotheses about the difference between two independent means;
- Use IBM SPSS to calculate t;
- Calculate effect size and estimate power;
- Understand the assumptions underlying the testing of hypotheses about the difference between two independent means;
- Understand the difference between random sampling and random assignment; and
- Appreciate how random sampling and random assignment serve as experimental controls.

In many fields, researchers often ask whether it makes a difference when some variable is measured under two different conditions. Is the reading comprehension of dyslexic children the same under conditions of normal and reduced visual contrast? Do AIDS patients who are given the drug AZT live longer than patients who are not given the drug? Is the error rate of typists the same when work is done in a noisy environment as in a quiet one?

Let us designate scores obtained under one condition as X and scores obtained under the second condition as Y. Now suppose that we draw one sample at random from the population of X scores and draw another sample at random from the population of Y scores. The two random samples are selected *independently* from their respective populations. **Independent samples** exist when the selection of elements comprising the sample of Y scores is in no way influenced by the selection of elements making up the sample of X scores, and vice versa.

independent samples samples drawn so that the selection of elements in one sample is in no way influenced by the selection of elements in the other, and vice versa

14.1 The Null and Alternative Hypotheses

Each of the questions posed in the introduction is an example of a problem for which we can use hypothesis testing to find the answer. However, the hypotheses you will test in this chapter are different from those you tested in Chapter 12. Here we are concerned with the hypothesis that *two* means are equal.

Consider the first question. Dyslexia refers to a specific reading disability that affects 10% to 15% of the general population. A dyslexic child is defined as a child of normal or above-normal intelligence with no known organic or gross behavioral disorders who, despite normal schooling, shows a reading lag of at least 2 years. Most intervention programs have focused on remediation of language skills, but with little success. Psychologist Mary Williams proposed that dyslexia was the result of a complex visual deficit that could be immediately corrected by simply putting a blue plastic overlay on the reading material. (The following procedure and results are modified, with permission, from Williams, LeCluyse, & Rock-Faucheux, 1992.) To test this, 24 dyslexic children were randomly assigned to one of two groups, one of which read from pages with a clear plastic overlay and the other of which read from pages with a blue plastic overlay. The dependent variable was percentage correct on a multiple-choice reading comprehension test. The mean of the group using clear overlays was 56.7%, whereas the mean of the group using the blue overlay was 75.0%. Is this a meaningful difference? *Because of the variability introduced by random assignment, we would expect two samples to have different means even if the two groups had been treated alike.*

The important question is not about the samples, but about the populations from which the samples came. Is it possible that the mean of the population of reading comprehension scores obtained with blue overlays is the same as the mean of the population of observations obtained with clear overlays? To answer this, we express the question as a statistical null hypothesis. If scores obtained under the blue overlay condition are designated by X and those obtained under the clear overlay condition by Y, we have:

$$H_0\text{: } \mu_X - \mu_Y = 0$$

The test of the difference between two means always involves the difference between measures of the *same* dependent variable taken under two conditions. In the dyslexic-children example, X stands for the reading comprehension scores of subjects treated one way, and Y stands for the reading comprehension scores of subjects treated another way.

The alternative hypothesis, H_A, may be directional or nondirectional, depending on the logic of the study. Williams and her colleagues wanted to know whether an effect exists in *either* direction and therefore chose a nondirectional alternative hypothesis:

$$H_0\text{: } \mu_X - \mu_Y = 0$$
$$H_A\text{: } \mu_X - \mu_Y \neq 0$$

From this point on, the procedure for testing the null hypothesis follows the same basic principles you learned in Chapter 12. We must choose α, the criterion for decision, and select a sample size that will give us a good chance of discovering a difference (if one exists). Next, we must examine the characteristics of the pertinent sampling distribution so that the present sample outcome can be compared with those outcomes expected when the null hypothesis is true. In the sampling distribution, we establish a region of rejection according to the nature of H_A and the value of α. In the present problem, if the difference between the two sample means is so great in either direction that it could not reasonably be accounted for by chance variation when the population means are the same, Williams and her colleagues will reject the null hypothesis.

14.2 The Random Sampling Distribution of the Difference between Two Sample Means

Recall from Chapter 12 that the random sampling distribution of *single* means described all the sample means that can possibly occur when sampling at random from a population having the mean specified in H_0. Here we are concerned with the *difference* between two population means, and the distribution that we will refer to is the random sampling distribution of the *difference* between two sample means. Just what is this distribution?

random sampling distribution of $(\overline{X} - \overline{Y})$ a theoretical relative frequency distribution of all the values of $(\overline{X} - \overline{Y})$ that would be obtained by chance from an infinite number of samples of a particular size drawn from two populations, μ_X and μ_Y

$\mu_{\overline{X}-\overline{Y}}$ symbol for the mean of the sampling distribution of $(\overline{X} - \overline{Y})$

We began this chapter by drawing one sample at random from the population of X scores and another from the population of Y scores. Now we compute the mean of each sample and obtain the difference between these two means. Suppose, then, that we return the samples to their respective populations and select a second pair of samples of the same size in the same way. We again calculate the two means and compute and record the difference between them. If we repeat this experiment indefinitely, the differences—the values of $(\overline{X} - \overline{Y})$ thus generated—form the **random sampling distribution of the difference between two sample means**.

This sampling distribution is very useful because it describes all the differences that would occur by chance between $\overline{X}$ and $\overline{Y}$ *when* H_0 *is true.* When testing for a difference between two means, H_0 is almost always $\mu_X - \mu_Y = 0$. Now if we draw pairs of samples at random from two populations, *and if the means of the two populations are the same*, because of chance factors we will find that sometimes $\overline{X}$ is larger than $\overline{Y}$—leading to a positive value for $(\overline{X} - \overline{Y})$—and sometimes it is the other way around—leading to a negative value for $(\overline{X} - \overline{Y})$. When all the possible differences are considered, their mean will be zero. Thus, *the mean of the sampling distribution of differences* $(\overline{X} - \overline{Y})$ *is zero when the difference between the two population means is zero.* We will use **$\mu_{\overline{X}-\overline{Y}}$** as the symbol for the mean of the sampling distribution of $(\overline{X} - \overline{Y})$. To summarize, then,

$$\mu_{\overline{X}-\overline{Y}} = 0$$

when $\mu_X - \mu_Y = 0$.

As an example of the sampling distribution of $(\overline{X} - \overline{Y})$, let us examine miniature populations. Suppose that the population of X scores consists of three scores: 3, 5, and 7. Because the question we usually ask in statistics is what would happen if the "treatment" had no effect, we will also suppose that the population of Y scores consists of the same three scores: 3, 5, and 7. If we draw samples of two cases with replacement from either population, there are nine possible samples and therefore nine possible means that may occur. These are shown in Table 14.1. Notice that we say that there are nine possible means, even though some of them have the same value. We must take all of them into account.

When we ask what differences may occur between sample means in X and sample means in Y, we must consider that there are nine sample means in X, and that each may be coupled

TABLE 14.1 *Possible Samples and Sample Means for Samples of Size 2 (Sampling with Replacement)*

POPULATION OF *X* SCORES: 3, 5, 7 **POPULATION OF *Y* SCORES 3, 5, 7**

$\mu_X = \mu_Y = 5$
$\sigma_X = \sigma_Y = 1.633$

Sample from X population	$\overline{X}$	*Sample from Y population*	$\overline{Y}$
3, 3	3	3, 3	3
3, 5	4	3, 5	4
3, 7	5	3, 7	5
5, 3	4	5, 3	4
5, 5	5	5, 5	5
5, 7	6	5, 7	6
7, 3	5	7, 3	5
7, 5	6	7, 5	6
7, 7	7	7, 7	7

with any one of the nine sample means in *Y*. Thus, there are 81 possible pairings of the means of *X* with the means of *Y*. The possible sample means are shown at the top and in the left column of Table 14.2, and the 81 differences between pairs $(\overline{X} - \overline{Y})$ are shown in the body of the table. *If the samples are drawn independently and at random, each of the listed differences is equally likely to appear.* If we collect these differences in a frequency distribution, the result is as shown in Figure 14.1.

Look at the resulting frequency distribution. Notice that the mean of the distribution (the expected value) is zero, the difference between μ_X and μ_Y. Notice also that the distribution is very reasonably approximated by the normal curve. *If the sampling distributions of the mean for* X *and* Y *are approximately normal, the distribution of the difference between pairs of means selected at random will also be approximately normal.* In the present situation, the distribution of the population of scores in *X* and *Y* resembles a rectangular distribution more than it does a normal distribution. Yet the distribution of the difference between pairs of sample means resembles the normal curve. Thus, *the effect predicted for single means by the central limit theorem* (see Section 11.4) *also applies to the distribution of the difference between sample means.*

TABLE 14.2 *Differences between Pairs of Sample Means* $(\overline{X} - \overline{Y})$ *for All Possible Paired Samples of* X *and* Y *(Data from Table 14.1)*[*]

					$\overline{Y}$				
$\overline{X}$	3	4	4	5	5	5	6	6	7
7	4	3	3	2	2	2	1	1	0
6	3	2	2	1	1	1	0	0	−1
6	3	2	2	1	1	1	0	0	−1
5	2	1	1	0	0	0	−1	−1	−2
5	2	1	1	0	0	0	−1	−1	−2
5	2	1	1	0	0	0	−1	−1	−2
4	1	0	0	−1	−1	−1	−2	−2	−3
4	1	0	0	−1	−1	−1	−2	−2	−3
3	0	−1	−1	−2	−2	−2	−3	−3	−4

[*] The body of the table contains values of $(\overline{X} - \overline{Y})$.

$(\overline{X}-\overline{Y})$	f
4	1
3	4
2	10
1	16
0	19
−1	16
−2	10
−3	4
−4	1
	$n = 81$

$\mu_{\overline{X}-\overline{Y}} = 0$

$\sigma_{\overline{X}-\overline{Y}} = 1.633$

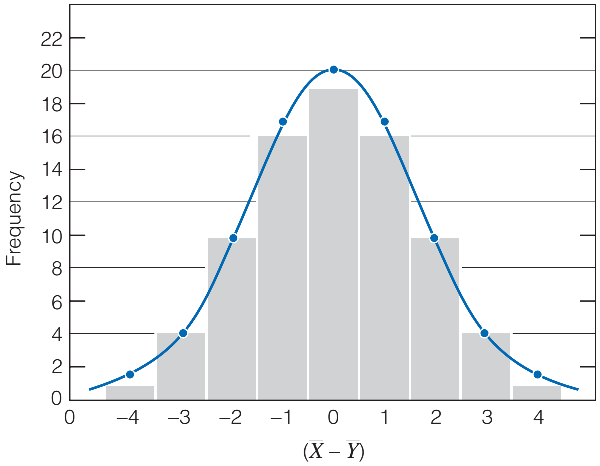


FIGURE 14.1 Distribution of differences between all possible pairs of sample means.

14.3 Properties of the Sampling Distribution of the Difference between Means

As we know, three characteristics completely define any distribution: mean, standard deviation, and shape. We now know that the mean of the random sampling distribution of the difference between pairs of sample means, $\mu_{\overline{X}-\overline{Y}}$, is the same as the difference between the two population means, $\mu_X - \mu_Y$. We also know that the sampling distribution of $(\overline{X}-\overline{Y})$ will be normally distributed when the two populations are normally distributed and that the sampling distribution tends toward normal even when the two populations are not normal. But what about its standard deviation?

The standard deviation of the sampling distribution of $(\overline{X}-\overline{Y})$ is called the **standard error of the difference between two means**, and its symbol is $\boldsymbol{\sigma_{\overline{X}-\overline{Y}}}$. If the sampling distribution is composed of differences between means of *independent* samples, its standard deviation is:

standard error of the difference between two means
the standard deviation of the sampling distribution of $(\overline{X}-\overline{Y})$

$\boldsymbol{\sigma_{\overline{X}-\overline{Y}}}$
symbol for the standard error of the sampling distribution of $(\overline{X}-\overline{Y})$

STANDARD ERROR OF THE DIFFERENCE
BETWEEN TWO INDEPENDENT MEANS

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\sigma^2_{\overline{X}} + \sigma^2_{\overline{Y}}} \qquad (14.1)$$

In the example of Section 14.2, the distribution of possible differences was constructed on the assumption of independence of sample means. According to Formula 14.1, its standard deviation should be

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\left(\frac{1.633}{\sqrt{2}}\right)^2 + \left(\frac{1.633}{\sqrt{2}}\right)^2} = 1.633$$

Notice that actual calculation of the standard deviation (see Figure 14.1) yields the same value, 1.633. It is just a coincidence that in this problem, $\sigma_X = \sigma_Y = \sigma_{\overline{X}-\overline{Y}}$.

14.4 Determining a Formula for *t*

In this section, we will develop a formula for *t* for testing the difference between the means of two independent groups. For those of you who are strong in mathematics, you will find the progression to be very logical. If you are not, we do not want this section to discourage you. If you want, read the next paragraph of this section (through Formula 14.2*a*) and then skip ahead to Formula 14.4.

Formula 14.1 requires that we know the standard error of the mean of X and of Y, and these, in turn, require that we know σ_X and σ_Y. However, we usually do not know the population standard deviations and so must estimate them from the samples. The symbol for the **estimated standard error of the difference between two means** is $s_{\overline{X}-\overline{Y}}$. The formula for estimating $\sigma_{\overline{X}-\overline{Y}}$ is:

estimated standard error of the difference between two means an estimate of the standard deviation of the sampling distribution of $(\overline{X} - \overline{Y})$

ESTIMATE OF $\sigma_{\overline{X}-\overline{Y}}$ FOR INDEPENDENT SAMPLES

$$s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2} \tag{14.2a}$$

To test the difference between two independent means, we could calculate an "approximate z" where $z = (score - mean)/standard\ deviation$. In the sampling distribution of differences, our difference—$(\overline{X} - \overline{Y})$—is the score, the hypothesized difference is the mean, and $s_{\overline{X}-\overline{Y}}$ is the estimated standard deviation:

$$\text{“}z\text{”} = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{s_{\overline{X}-\overline{Y}}}$$

When the sample size is relatively large, "z" is very nearly normally distributed. As sample size decreases, its distribution departs from the normal, but, unfortunately, *except in the special case when* $n_X = n_Y$, neither is it distributed exactly as Student's *t*.

The eminent British statistician Sir Ronald A. Fisher showed that a slightly different approach results in a statistic that *is* distributed as Student's *t*. The change that Fisher introduced was, in effect, to assume that $\sigma_X^2 = \sigma_Y^2$. This is often called the **assumption of homogeneity of variance**. To understand the consequences of this assumption, it will be helpful to rewrite Formula 14.1 for $\sigma_{\overline{X}-\overline{Y}}$. According to Formula 11.2, we may substitute σ_X^2/n_X for $\sigma_{\overline{X}}^2$ and σ_Y^2/n_Y for $\sigma_{\overline{Y}}^2$. The result is

assumption of homogeneity of variance the assumption that $\sigma_X^2 = \sigma_Y^2$, that the variances in the two populations from which the samples are drawn are the same

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

This formula clearly allows for the possibility that σ_X^2 might be different from σ_Y^2 and so does the following formula, which is used when *estimates* of the two variances must be made from the *samples*:

$$s_{\overline{X}-\overline{Y}} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} \tag{14.2b}$$

If σ_X^2 and σ_Y^2 differ, then each sample estimates its corresponding population value. Under the assumption of homogeneity of variance, however, both s_X^2 and s_Y^2 are estimates of the same population variance. If this is so, then rather than make two separate estimates, each based on a small sample, it is preferable to combine the information from both samples and make a single **pooled estimate of the population variance** (symbolized s_p^2). In general, the "best estimate" of a population variance is made as follows:

pooled estimate of the population variance (s_p^2) a single estimate of the population variance based on combining the data from both samples

$$\sigma^2_{\text{est}} = \frac{\textit{sum of squares of deviation scores}}{\textit{degrees of freedom}} = \frac{SS}{df}$$

This is exactly what we do when we calculate

$$s_X^2 = \frac{\sum(X - \overline{X})^2}{n - 1}$$

We can make a pooled estimate from the two samples by pooling the sums of squares of the deviation scores from each sample and dividing by the total number of degrees of freedom associated with the pooled sum of squares. Thus, s_p^2 is:

$(n_X - 1) + (n_Y - 1)$ degrees of freedom for the pooled estimate of the population variance obtained from two independent samples

ESTIMATE OF σ^2 MADE BY POOLING INFORMATION FROM TWO SAMPLES

$$s_p^2 = \frac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)} \tag{14.3}$$

$s_{\overline{X}-\overline{Y}}$ symbol for the estimated standard error of $(\overline{X} - \overline{Y})$

We may substitute this quantity for s_X^2 and for s_Y^2 in the formula for $s_{\overline{X}-\overline{Y}}$:

$$s_{\overline{X}-\overline{Y}} = \sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}$$

Factoring out s_p^2, the formula may be rewritten as

$$s_{\overline{X}-\overline{Y}} = \sqrt{s_p^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$$

$\sqrt{\frac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$ general formula for the estimated standard error of $(\overline{X} - \overline{Y})$ in the independent-samples design

If the expression for s_p^2 given by Formula 14.3 is substituted in this formula, we have

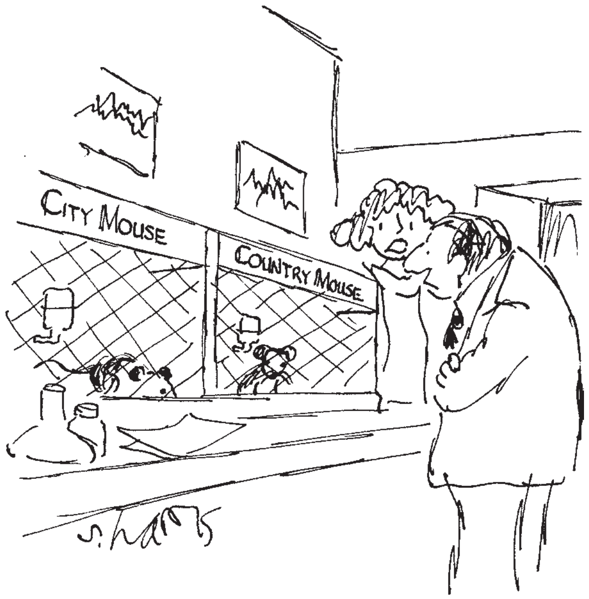
ESTIMATE OF STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO INDEPENDENT MEANS WHEN $\sigma_X^2 = \sigma_Y^2$

$$s_{\overline{X}-\overline{Y}} = \sqrt{\frac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)} \tag{14.4}$$

When the two *n*s are equal, we can calculate $s_{\overline{X}-\overline{Y}}$ by a simpler formula than Formula 14.4. It is:

ESTIMATE OF STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO INDEPENDENT MEANS WHEN $\sigma_X^2 = \sigma_Y^2$ AND $n_X = n_Y = n$

$$s_{\overline{X}-\overline{Y}} = \sqrt{\frac{SS_X + SS_Y}{n(n - 1)}} \tag{14.5}$$



"WE PLAN TO DETERMINE, ONCE AND FOR ALL, IF THERE REALLY ARE ANY CULTURAL DIFFERENCES BETWEEN THEM."

© Sidney Harris, ScienceCartoonsPlus.com

$t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{s_{\overline{X}-\overline{Y}}}$

formula for the t statistic for testing a null hypothesis about the difference between two population means

When we calculate t according to the following formula and calculate $s_{\overline{X}-\overline{Y}}$ according to Formula 14.4 or 14.5, t will be distributed according to Student's distribution with $(n_X - 1) + (n_Y - 1)$ degrees of freedom:

t FOR TESTING HYPOTHESES ABOUT
THE DIFFERENCE BETWEEN TWO MEANS

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{s_{\overline{X}-\overline{Y}}} \qquad (14.6)$$

14.5 Testing the Hypothesis of No Difference between Two Independent Means: The Dyslexic Children Experiment

As an example of how to use Formula 14.6, let us return to the study we introduced at the beginning of the chapter. Is the reading ability of dyslexic children improved by placing a blue plastic overlay on reading material? Williams and her colleagues (1992) tested the statistical hypothesis that

$$H_0\colon \mu_X - \mu_Y = 0$$

$$H_A\colon \mu_X - \mu_Y \neq 0$$

using the 5% level of significance ($\alpha = .05$). Twenty-four dyslexic children were randomly assigned to one of the two groups. The data for this problem and the calculations are shown in Table 14.3.

TABLE 14.3 *Calculations for Testing a Hypothesis about the Difference between Two Independent Means for the Dyslexic-Children Experiment*

BLUE OVERLAY X	CLEAR OVERLAY Y	X^2	Y^2
70	50	4,900	2,500
80	40	6,400	1,600
90	50	8,100	2,500
80	50	6,400	2,500
50	60	2,500	3,600
80	60	6,400	3,600
70	60	4,900	3,600
80	40	6,400	1,600
70	60	4,900	3,600
80	70	6,400	4,900
80	60	6,400	3,600
70	80	4,900	6,400
$\sum X = 900$	$\sum Y = 680$	$\sum X^2 = 68{,}600$	$\sum Y^2 = 40{,}000$

① $\overline{X} = \dfrac{\sum X}{n_X} = \dfrac{900}{12} = 75.0 \qquad \overline{Y} = \dfrac{\sum Y}{n_Y} = \dfrac{680}{12} = 56.7$

②a $SS_X = \sum X^2 - \dfrac{\left(\sum X\right)^2}{n_X} = 68{,}600 - \dfrac{900^2}{12} = 1{,}100$

$$s_X = \sqrt{\frac{SS_X}{n_X - 1}} = \sqrt{\frac{1{,}100}{12 - 1}} = \sqrt{100} = 10$$

②b $SS_Y = \sum Y^2 - \dfrac{\left(\sum Y\right)^2}{n_Y} = 40{,}000 - \dfrac{680^2}{12} = 1466.7$

$$s_Y = \sqrt{\frac{SS_Y}{n_Y - 1}} = \sqrt{\frac{1466.7}{12 - 1}} = \sqrt{133.3} = 11.5$$

③ $t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{\sqrt{\left(\dfrac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\right)\left(\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right)}}$

$$= \frac{(75.0 - 56.7) - 0}{\sqrt{\left(\dfrac{1{,}100 + 1466.7}{(12 - 1) + (12 - 1)}\right)\left(\dfrac{1}{12} + \dfrac{1}{12}\right)}}$$

$= +4.15$

④ $df = (n_X - 1) + (n_Y - 1) = (12 - 1) + (12 - 1) = 22$

⑤ $t_{\text{crit}} = \pm 2.074$

Decision: Reject H_0

We begin our calculations by finding n_X, ΣX, ΣX^2, n_Y, ΣY, and ΣY^2. In this example, there were 12 scores under each condition, but we nevertheless used Formula 14.4 to calculate t. Because $n_X = n_Y$, we could have used Formula 14.5, but it seemed best to demonstrate the more general procedure, which is appropriate when $n_X \neq n_Y$ as well as when both are the same.

Next, we calculate the two means in Step ① and the two sums of squares of the deviation scores in Steps (2a) and (2b). Although the two estimated standard deviations are not required for the subsequent calculations, good practice requires reporting them, and the necessary calculations are shown. Step ③ shows the calculation of t, and Step ④ shows the determination of the degrees of freedom. Notice that when we test null hypotheses of no difference between two independent means, the number of degrees of freedom is $(n_X - 1) + (n_Y - 1)$. In Step ⑤, the critical value of t is obtained from Table D in Appendix D.

The outcome of the test is shown in Figure 14.2. The value of t obtained by Williams and her colleagues falls in the region of rejection. It is so different from the value stated in H_0 that

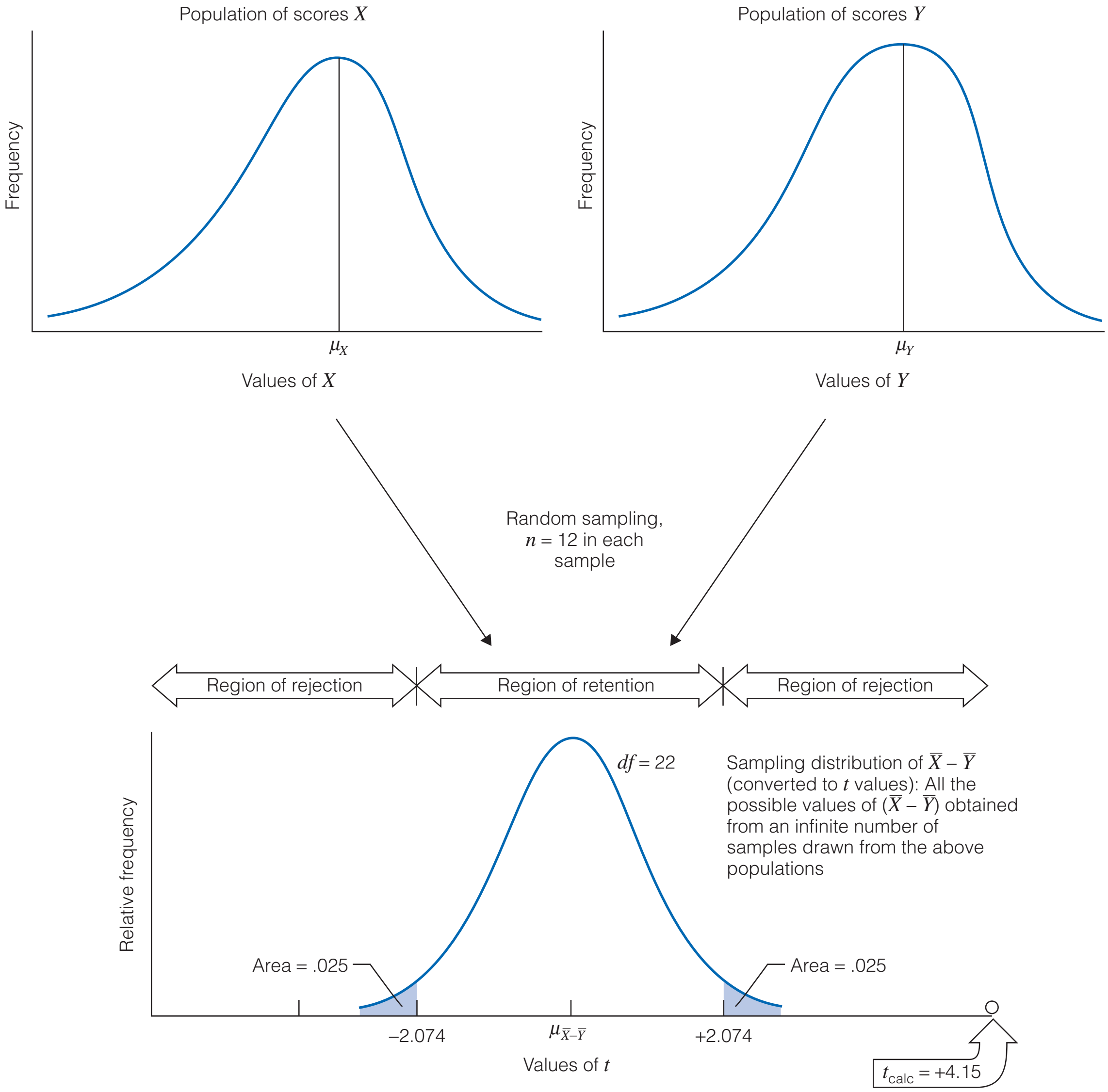

Population of scores X
Frequency
μ_X
Values of X
Population of scores Y
Frequency
μ_Y
Values of Y
Random sampling, n = 12 in each sample
Region of rejection
Region of retention
Region of rejection
Relative frequency
df = 22
Sampling distribution of X̄ – Ȳ (converted to t values): All the possible values of (X̄ – Ȳ) obtained from an infinite number of samples drawn from the above populations
Area = .025
Area = .025
−2.074
μ_X̄–Ȳ
+2.074
Values of t
t_calc = +4.15

FIGURE 14.2 Testing a hypothesis about the difference between two means against a nondirectional alternative; $\alpha = .05$.

they concluded that it had a low probability of occurring by chance if H_0 is true. Thus, they rejected H_0. Their research conclusion was that blue plastic overlays on reading material improve visual processing and thereby provide immediate improvement in the reading ability of dyslexic children.

14.6 Use of a One-Tailed Test

directional (one-tailed) test the alternative hypothesis, H_A, states that μ_X differs from μ_Y in one particular direction (and the critical region is located in only one tail of the sampling distribution)

Suppose it had been important in the dyslexic-children study to discover only whether the blue overlay *improved* reading. In this case, Williams and colleagues would have used a one-tailed test. The null hypothesis and the alternative hypothesis would then have been

$$H_0: \mu_X - \mu_Y = 0$$

$$H_A: \mu_X - \mu_Y > 0$$

We must now place the region of rejection entirely in the upper tail of the sampling distribution of $(\overline{X} - \overline{Y})$. In the t distribution with degrees of freedom = 22, $t = +1.717$ identifies the point above which 5% of the values will be found. Figure 14.3 illustrates the situation. Our calculation of t is exactly the same as for a two-tailed test, and the obtained value remains $t = +4.15$. Our obtained t again falls in the region of rejection.

14.7 Calculation of *t* with IBM SPSS

In this section, you will learn how to use IBM SPSS to calculate the t statistic to test hypotheses about two independent means. You will use the data in Table 14.3.

It is important to note that the scores from the two groups are on the same dependent variable taken under two different conditions (i.e., blue overlay versus clear overlay). In the **Data Editor**, all the scores should be entered in a single column. Now, in the next column, we need to identify from what condition the reading comprehension scores came. We will enter 1s for scores of subjects in the blue overlay condition and 2s for scores of subjects in the clear overlay condition. In the **Variable View**, for our dependent variable, we will use Scores as the variable name and Reading Comprehension Scores as the label. For the second variable that was used to indicate group membership, we will use Condition for the variable name and Type of Overlay as the label.

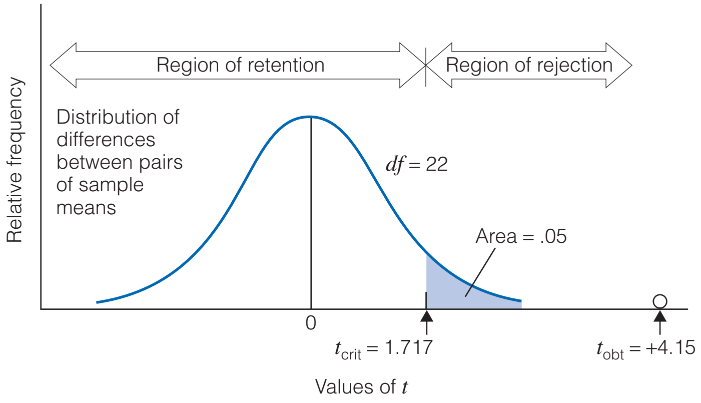
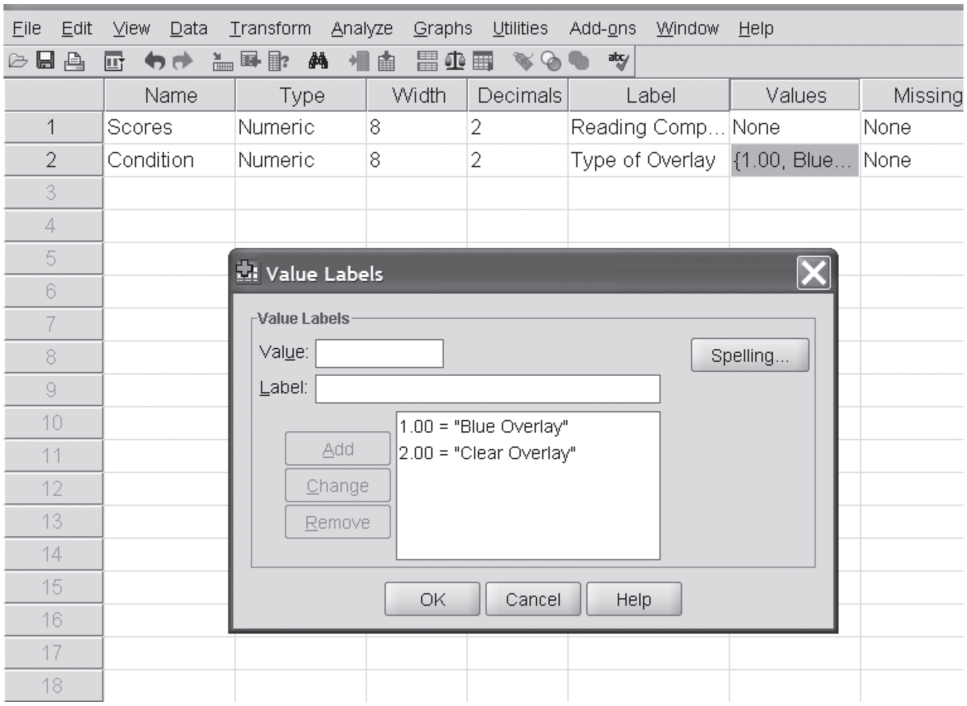


FIGURE 14.3 Testing the hypothesis that $\mu_X - \mu_Y = 0$ against the alternative that $\mu_X - \mu_Y > 0$; data from Table 14.3.



	Name	Type	Width	Decimals	Label	Values	Missing
1	Scores	Numeric	8	2	Reading Comp...	None	None
2	Condition	Numeric	8	2	Type of Overlay	{1.00, Blue...	None
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

Value Labels

Value Labels

Value:

Label:

1.00 = "Blue Overlay"
2.00 = "Clear Overlay"

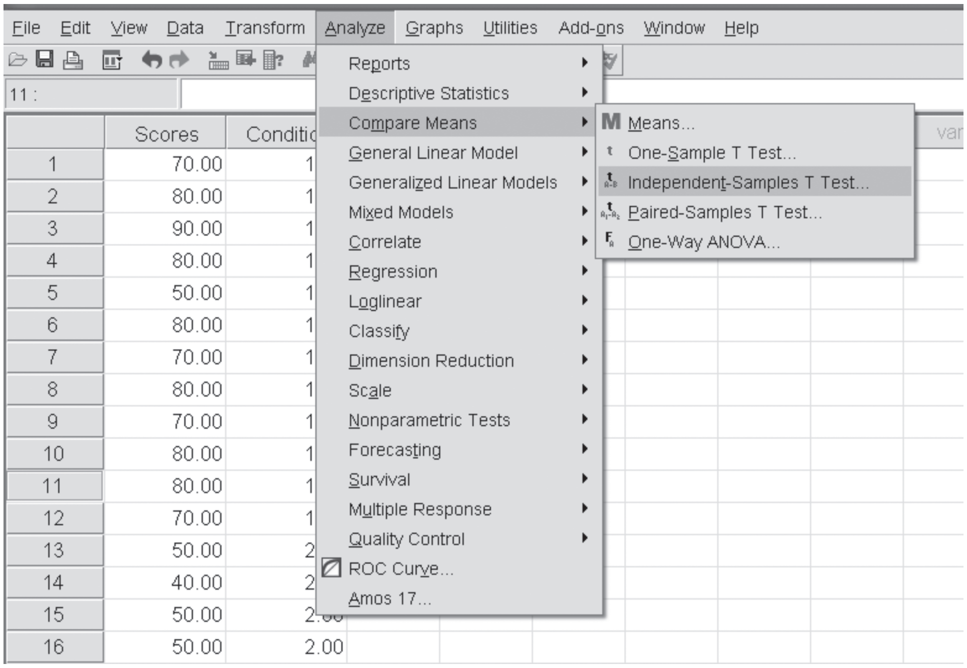
FIGURE 14.4 Value Labels dialog box in IBM SPSS.

Now that the data, variable names, and labels have been entered, we will use a feature in IBM SPSS that we have not yet introduced. In the **Variable View**, it is the column labeled Values. It is useful for instances when data can be placed into different groups. For example, in our data, the variable called Condition is used to identify groupings in our data. Instead of referring to the groups as condition 1 and condition 2, we can assign a descriptive label to the 1s and 2s. Now, for the variable called Condition, click on Values. A dialog box called Value Labels will appear. In the Value box, type "1" and, in the Label box, type "Blue Overlay." Click Add. Now, in the Value box, type "2" and, in the Label box, type "Clear Overlay." Click Add. Figure 14.4 shows the dialog box after the values and their labels have been correctly added. Click OK.

To conduct a hypothesis test on two independent means in IBM SPSS, click **Analyze**, then select **Compare Means** followed by **Independent-Samples T Test.** Figure 14.5 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze → Compare Means → Independent-Samples T Test.**

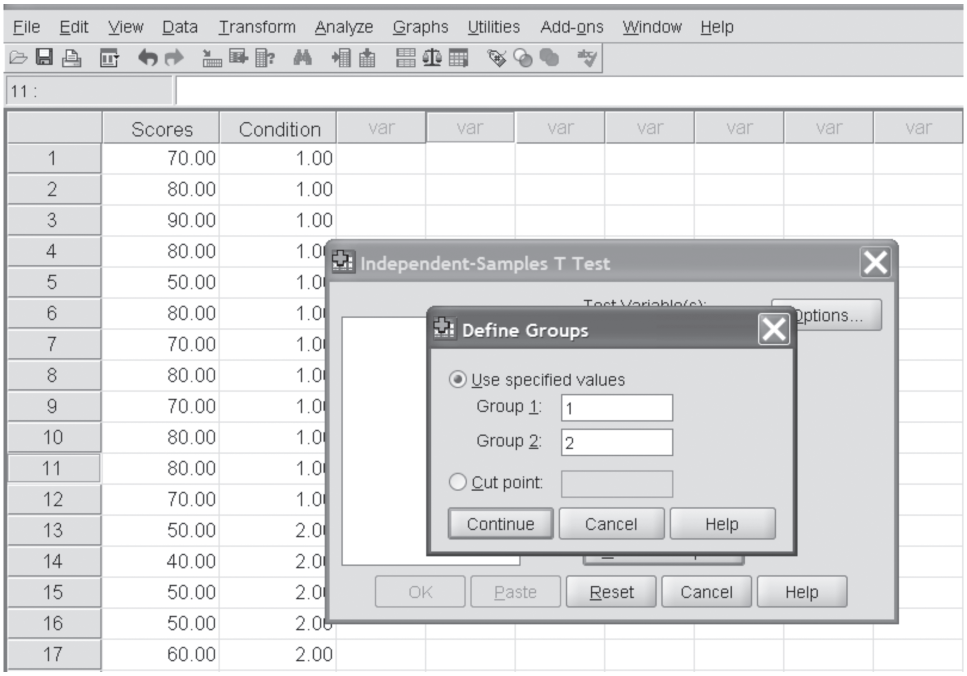
Now, a dialog box will appear. Move the Reading Comprehension Scores variable (i.e., Scores) to the Test Variable(s) box and the Type of Overlay variable (i.e., Condition) to the Grouping Variable box. Just below the Grouping Variable box, click the Define Groups button. Here, you will specify the two groups to be compared. For Group 1, type "1" and for Group 2, type "2" since we used 1s and 2s in the **Data Editor** to identify the different groups. Figure 14.6 shows the dialog box after you have correctly specified the groups to be compared. Click Continue, then OK.

Table 14.4 shows the IBM SPSS output. After the syntax, there is a section of output labeled Group Statistics. As you can see, it contains various statistics, but for each condition. For example, we see that the sample size in the blue overlay and clear overlay conditions was 12 in each group. In addition, we see that the mean in the blue overlay and


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
11 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
var

	Scores	Conditio
1	70.00	1
2	80.00	1
3	90.00	1
4	80.00	1
5	50.00	1
6	80.00	1
7	70.00	1
8	80.00	1
9	70.00	1
10	80.00	1
11	80.00	1
12	70.00	1
13	50.00	2
14	40.00	2
15	50.00	2
16	50.00	2.00

FIGURE 14.5 IBM SPSS menus when calculating two independent samples *t*.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
11 :
var
Independent-Samples T Test
Test Variable(s):
Options...
Define Groups
Continue
Cancel
Help
OK
Paste
Reset
Cancel
Help

	Scores	Condition
1	70.00	1.00
2	80.00	1.00
3	90.00	1.00
4	80.00	1.0
5	50.00	1.0
6	80.00	1.0
7	70.00	1.0
8	80.00	1.0
9	70.00	1.0
10	80.00	1.0
11	80.00	1.0
12	70.00	1.0
13	50.00	2.0
14	40.00	2.0
15	50.00	2.0
16	50.00	2.0
17	60.00	2.00

● Use specified values
Group 1: 1
Group 2: 2
○ Cut point:

FIGURE 14.6 IBM SPSS dialog box for defining groups when calculating two independent samples *t*.

TABLE 14.4 *IBM SPSS Output from Two Independent Samples t*

```
T-TEST GROUPS=Condition (1 2)
  /MISSING=ANALYSIS
  /VARIABLES=Scores
  /CRITERIA=CI (.95).
```

Group Statistics

	Type of Overlay	N	Mean	Std. Deviation	Std. Error Mean
Reading Comprehension Scores	Blue Overlay	12	75.0000	10.00000	2.88675
	Clear Overlay	12	56.6667	11.54701	3.33333

Independent Samples Test

		Levene's Test for Equality of Variances		*t*-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	*t*	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Reading Comprehension Scores	Equal variances assumed	270	.609	4.158	22	.000	18.33333	4.40959	9.18841	27.47825
	Equal variances not assumed			4.158	21.560	.000	18.33333	4.40959	9.17758	27.48909

clear overlay conditions, respectively, was 75 and 56.6667. IBM SPSS also provides the estimated standard deviation, and estimated standard error of the mean in each condition. The next section of output is labeled Independent Samples Test. You will notice that the results are presented in two rows: one labeled Equal variances assumed and a second labeled Equal variances not assumed. We will not focus on the second row, which has to do with instances when the assumption of homogeneity of variance (see Section 14.4) is violated. Instead, we will discuss the first row, which deals with the case when the assumption of homogeneity of variance is not violated. The column labeled Levene's Test for Equality of Variances provides a formal statistical test of the homogeneity of variance assumption. Although it is beyond the scope of this book, we refer the interested reader to Fox (2008) for a discussion of it and related procedures. Next, we see that the t is 4.158 and the $df = 22$. Notice that the two-tailed p-value associated with the t is a probability less than .001. You should not be interpreting the Sig. (2-tailed) as $p = 0$. IBM SPSS only displays the first three numbers to the right of the decimal. However, if you would like to see the actual p-value, you can double-click on the table and select the cell under Sig. (2-tailed). The next two columns present the mean difference (i.e., $75 - 56.6667$) and estimated standard error of the difference (see Formula 14.4), respectively. The last two columns will be discussed later in Section 17.7.

14.8 Sample Size in Inference about Two Means

effect size an estimate of the degree to which the treatment effect is present in the population, expressed as a "number...free of the original measurement unit" (Cohen, 1988)

When we test a hypothesis about two independent sample means, the samples do not have to be the same size. However, if σ_X and σ_Y are equal, when $n_X = n_Y$ the result is a smaller value of $\sigma_{\overline{X}-\overline{Y}}$ than otherwise. For example, when one sample is twice as large as the other, $\sigma_{\overline{X}-\overline{Y}}$ is about 6% larger than it would be if the same total sample size were distributed equally between the two samples (assuming $\sigma_{\overline{X}} = \sigma_{\overline{Y}}$). The inflation of $\sigma_{\overline{X}-\overline{Y}}$ continues at an accelerated pace as the relative sample sizes become increasingly more discrepant. The advantage of a smaller $\sigma_{\overline{X}-\overline{Y}}$ is that, if there *is* a difference between μ_X and μ_Y, the probability of rejecting H_0 is increased.

The preceding discussion relates to the *relative* sizes of the two samples. But what about the *absolute* magnitudes of the sample sizes? Other things being equal, large samples increase the probability of detecting a difference when a difference exists. We discuss the question of the optimum sample size for testing hypotheses about two means in Section 14.10.

14.9 Effect Size

(Cohen's) *d* a measure of effect size; expresses the difference between $(\mu_X - \mu_Y)_{\text{true}}$ and $(\mu_X - \mu_Y)_{\text{hyp}}$ relative to the population standard deviation

(Hedges') *g* a measure of effect size; expresses the difference between $\overline{X}$ and $\overline{Y}$ relative to s_p, the pooled estimate of the population standard deviation

Recall from Chapter 13 that it is often important to express the magnitude of the difference between μ_{hyp} and μ_{true} in terms of the number of standard deviations of the measure under consideration. We referred to such a measure as the effect size, *d* (Cohen, 1965, 1988). For tests of hypotheses about the difference between two independent means, *and assuming that* $\sigma_X = \sigma_Y = \sigma$:

$$d = \frac{(\mu_X - \mu_Y)_{\text{true}} - (\mu_X - \mu_Y)_{\text{hyp}}}{\sigma} \qquad (14.7)$$

Notice that *d* expresses the difference *relative to the population standard deviation of the set of measures*, not to the standard error.

To get a feel for effect sizes of different magnitudes, look at Figure 14.7. It shows the amount of overlap between two populations with means that differ by .25, .50, .75, and 1.0 standard deviations.

Ordinarily, σ is not known, so we must estimate it. In the case of two independent samples, we use the pooled estimate. Recall also that ordinarily, $(\mu_X - \mu_Y)_{\text{hyp}} = 0$. Therefore, once we have completed our study, we may estimate effect size by *g* (Hedges & Olkin, 1985):

FORMULA FOR EFFECT SIZE, *g*

$$\text{estimated } d = g = \frac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}}} \qquad (14.8)$$

where, similar to Cohen's *d* (1988; 1992), 0.2 = small effect, 0.5 = medium effect, and 0.8 = large effect.

Remember, effect size does not tell us the same thing as a *p*-value (Section 13.2). Highly significant *p*-values do not indicate that there is a large effect. Conversely, a nonsignificant result may still be important as judged by the effect size. As an example, consider the following two studies described by Rosnow and Rosenthal (1989a). (The studies are real, but the names

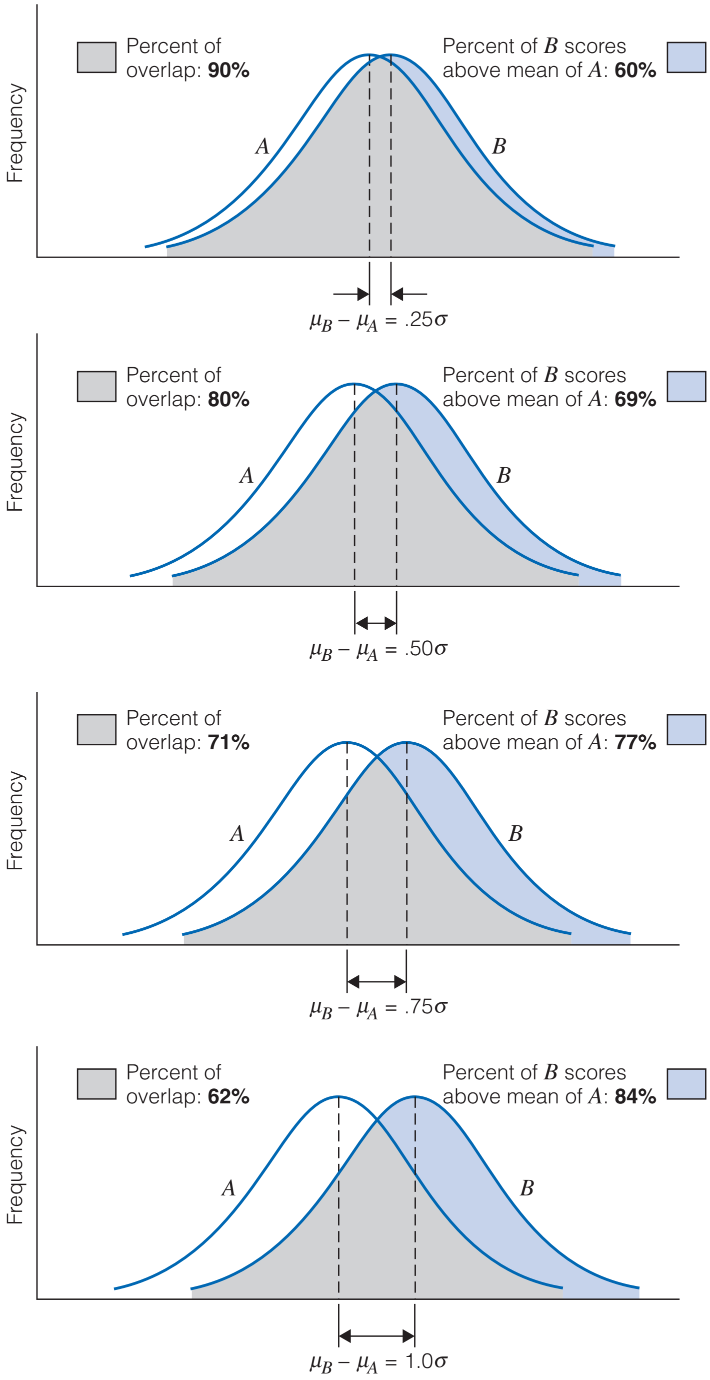


FIGURE 14.7 Amount of overlap of two normally distributed populations with means that differ by .25 σ (top), .50 σ, .75 σ, or 1.0 σ (bottom).

have been changed.) Professor Smith uses 80 subjects to compare two styles of leadership and discovers that style A is significantly better than style B at fostering productivity ($t = 2.21$, $df = 78$, $p < .05$). Professor Jones, who invented style B, is not pleased with this result and replicates the study using only 20 subjects; he reports nonsignificant results ($t = 1.06$, $df = 18$, $p > .30$). Although the *p*-values differ substantially, the estimated effect size is the same for both studies ($d = .50$). Thus, the second study did not really contradict the first. Professor Jones' power to reject the null hypothesis was much lower than Professor Smith's because of the smaller sample size. Professor Jones' p-value is not impressive, but his effect size indicates a difference that may have practical importance.

The two measures of effect size, d and g, are members of what is known as the d family of effect sizes. Many statisticians prefer to express effect size as a correlation coefficient (see Rosenthal & DiMatteo, 2001; Rosnow, Rosenthal, & Rubin, 2000). For testing a hypothesis about the difference between two independent means:

r
point-biserial between groups r; a correlational measure of effect size = $\sqrt{\frac{t^2}{t^2 + df}}$

FORMULA FOR EFFECT SIZE, r

$$r = \sqrt{\frac{t^2}{t^2 + df}} \qquad (14.9)$$

where, df = degrees of freedom for t ($n_X - 1 + n_Y - 1$) for the two-group design. Here, **r** is not Pearson's r; it is what is called a point-biserial correlation between groups (X and Y; experimental and control) and scores on a continuous variable.

Compared to d or g, r more directly "represents the relationship between two levels of the independent variable and scores on the dependent variable" (Rosenthal & DiMatteo, 2001). Recall from Section 9.5 that r^2, the coefficient of determination, gives the proportion of variance in one variable that is associated with differences in another variable.

As examples, let's return to the dyslexic-children experiment (Table 14.3). By Formula 14.8,

$$g = \frac{75.0 - 56.7}{\sqrt{\frac{1{,}100 + 1466.7}{(12 - 1) + (12 - 1)}}} = \frac{18.3}{10.8} = 1.69$$

By Formula 14.9,

$$r = \sqrt{\frac{(4.15)^2}{(4.15)^2 + 22}} = \sqrt{.439} = .66$$

So, based on the coefficient of determination, 0.44 of the variance in reading comprehension test scores is associated with differences in the color of the plastic overlay.

The two measures of effect size, g and r, can be transformed from one to the other (Rosnow, Rosenthal, & Rubin, 2000). For the two-sample, equal-n study:

FORMULA FOR CONVERTING
g TO r ($n_X = n_Y$)

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{df}{n_{\text{total}}}\right)}} \qquad (14.10)$$

Take the value we just calculated for g in the dyslexic-children experiment and convert it to r. It works! For unequal sample sizes, use the following formula:

FORMULA FOR CONVERTING
g TO r $(n_X \neq n_Y)$

$$r = \frac{g}{\sqrt{g^2 + 4\left(\frac{\bar{n}}{n_h}\right)\left(\frac{df}{n_{\text{total}}}\right)}} \tag{14.11}$$

where: $\bar{n}$ = arithmetic mean sample size

n_h = harmonic mean sample size

$$= \frac{2n_X n_Y}{n_X + n_Y}.$$

What about the sign of the effect size correlation coefficient? It has become a convention to report it as positive "when the effect is in the predicted direction and as negative when the effect is in the unpredicted direction" (Rosnow, Rosenthal, & Rubin, 2000).

There is a whole family of r effect size measures. In the chapters to come, we will introduce you to some other members of this family.

14.10 Estimating Power and Sample Size for Tests of Hypotheses about the Difference between Two Independent Means

Recall that we can determine the amount of power for a particular sample size or, conversely, the sample size required to achieve a desired level of power if we can specify the following:

1. The risk (α level) we are willing to take of rejecting a true null hypothesis.
2. The magnitude of the *smallest* difference that, if it exists, we wish to be reasonably sure of discovering.

power of a test the probability of rejecting a false null hypothesis

Let us adopt $\alpha = .05$ and test H_0: $\mu_1 - \mu_2 = 0$ using a two-tailed test with independent samples. (We assume that $\sigma_1 = \sigma_2$, or nearly so.) The relationships are shown in Figure 14.8. Effect size is on the horizontal axis and power is on the vertical axis. The power curves (as Cohen calls them) show the relationship between effect size and power for a variety of sample sizes. (The value for n is the number of cases in *each* group. If the samples differ in size but neither is more than twice the size of the other, we may use the average of the two sample sizes for n, and the power values will be sufficiently accurate for most purposes.) The power curves shown in Figure 14.7 apply only to the independent samples case.

Recall the experiment conducted by the Centers for Disease Control that compared deaths in AIDS patients given the drug AZT or a placebo (Fischl et al., 1987; see the introduction to Chapter 1). Suppose that AZT actually does result in fewer deaths and that the difference between the means of the AZT and placebo populations is $d = .7$ (a rather large difference). If the research team had used 15 subjects per group, what would the power of their test have been? To determine this, we locate .7 on the horizontal axis (the effect size) and move vertically up as shown by the dashed line to the curve for $n = 15$ and then horizontally to the vertical axis (power), where we find that power is about .45. This means that if the difference between μ_{AZT} and μ_{placebo} really is $d = .7$, the team could expect to reject the null hypothesis and conclude that there is a difference in only 45 out of 100 identical replications of the experiment; in the remaining 55 replications, chance would lead them to retain H_0. The sample size of 15 subjects per group does not give them a very good chance of detecting this large

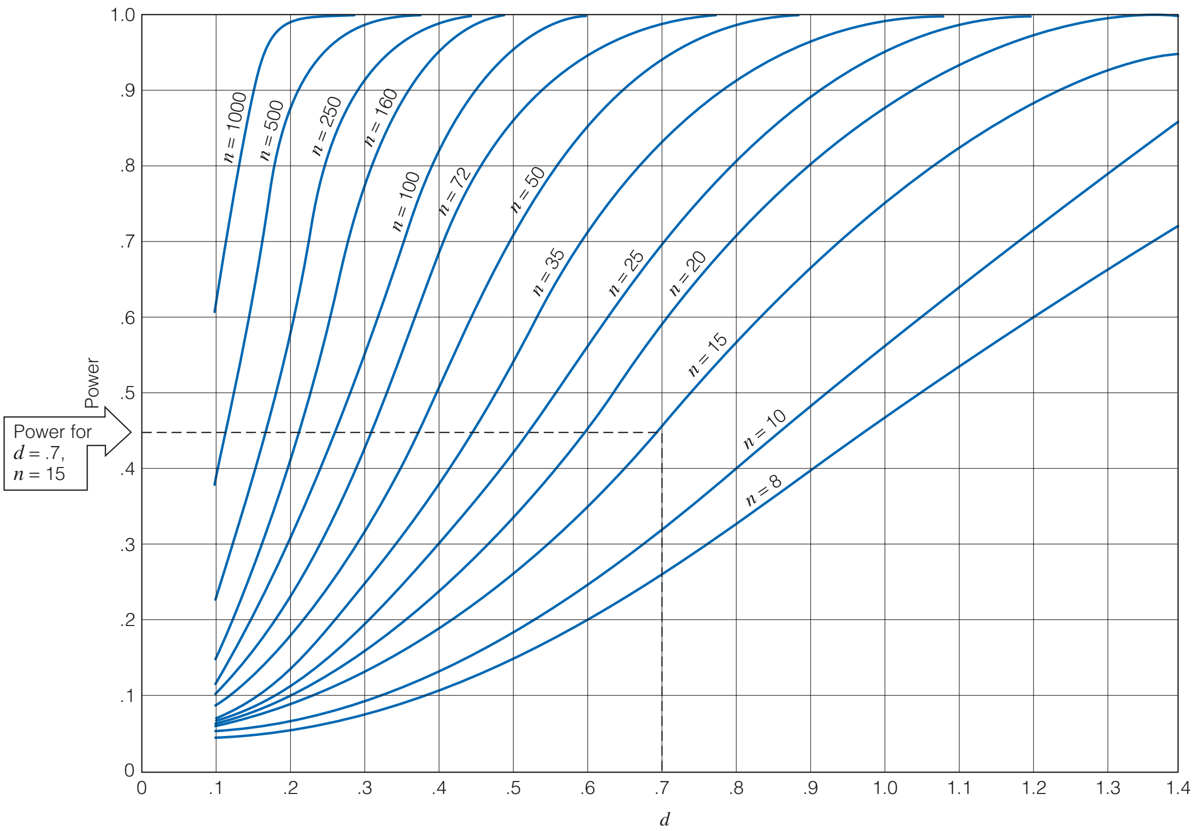


FIGURE 14.8 Power curves for the *t* test of H_0: $\mu_1 - \mu_2 = 0$ (independent samples; $\alpha = .05$) against a two-tailed alternative. (*Source*: Data for this figure were obtained from J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed., Lawrence Erlbaum Associates, New York, 1988. Used with permission.) Here, *n* is the number of cases in each group.

(and in this case, important) difference. Notice that for a smaller effect size, 15 subjects per group affords even less certainty of detecting a difference.

All too frequently, studies are performed by setting the level of significance and then deciding intuitively (or out of convenience) on sample size. Power curve charts such as Figure 14.7 make a more rational approach possible. Using the charts, one decides on a level of power and then selects sample size *according to the minimum effect size one wishes to discover* (if it exists). For example, if the CDC had wanted a probability of .9 of finding a medium-sized effect ($d = .5$) to be significant, a sample size of about 86 would have been required. In actuality, they used an *n* of about 140.

14.11 Assumptions Associated with Inference about the Difference between Two Independent Means

When we test a hypothesis about the difference between two means, the statistical conditions must be fundamentally the same as those previously described for the normal curve model (see Section 13.8). The major points are summarized as follows:

1. Each sample is drawn at random from its respective population.
2. The two random samples must be independently selected.

3. Samples are drawn with replacement.
4. The sampling distribution of $(\overline{X} - \overline{Y})$ follows the normal curve.

Under the normal curve model, the ideal condition requires that we know σ_X and σ_Y. Use of the *t* statistic and evaluation of the outcome in terms of Student's distribution of *t* evades this assumption. The *t* distribution makes allowance for the estimation of σ_X and σ_Y from the samples.

A second exception applies only to inference about the difference between two *independent* means. For the *t* distribution to be exactly applicable, homogeneity of variance is assumed. At first, this assumption sounds formidable. In practical application, there is help from several quarters.

First, experience suggests that the assumption of homogeneity of variance appears to be reasonably satisfied in many cases. It is relatively rare that a striking difference in variance occurs. Second, violation of the assumption causes less disturbance when samples are large than when they are small. As a rule of thumb, a moderate departure from the homogeneity of variance will have little effect when each sample consists of 20 or more observations. Finally, the problem created by heterogeneity of variance is minimized when the two samples are of equal (or nearly equal) size. There are other benefits from using equal sample sizes; look back at Section 14.8.

central limit theorem regardless of the shape of the population(s) of scores from which random samples are drawn, the sampling distribution of $\overline{X}$ (or $\overline{X} - \overline{Y}$) more and more resembles a normal distribution as sample size increases

Another point deserves special mention. A requirement, as you will remember, is that the sampling distribution of $(\overline{X} - \overline{Y})$ follows the normal distribution. Strictly speaking, the sampling distribution will be normal only when the distribution of the population of scores is also normal. However, according to the **central limit theorem**, the sampling distribution of the mean tends toward normality even when the population of scores is not normally distributed (see Section 11.4). *The strength of this tendency toward normality is pronounced when samples are large but is less so when samples are small.* Fortunately, the effect of the central limit theorem is rather pronounced unless sample size is quite small. For example, a moderate degree of skewness in the population can probably be tolerated if sample size is, say, 25 or more. When we have a serious question about the normality of the parent populations, a "nonparametric" statistical test will be better (see Chapter 21).

14.12 The Random-Sampling Model versus the Random-Assignment Model

random-sampling model a procedure that ensures that a sample is obtained in a way that all samples of the same size have an equal chance of being selected from the population

The goal of inferential statistics is to infer the nature of a population parameter from study of a sample. One of the basic assumptions that allows us to do that is random sampling from the population of interest. The **random-sampling model** makes it possible for us to know the characteristics of sampling distributions, which show the influence of chance sampling variation. We can then make *statistical conclusions* based directly on the laws of probability and properly generalize those conclusions to the entire population.

We discussed problems in selecting random samples and drawing proper conclusions in Section 13.12. (A random sample refers to a set of observations, but for convenience we will sometimes refer to a sample of subjects in this and the next two sections—see the introduction to Chapter 11.) Recall that in the real world, it is usually impossible for investigators to work with samples that are randomly drawn from the population of interest. Most work with convenient samples. For example, biomedical researchers rely on volunteers for subjects. Faculty members who do research with humans often use students who are currently enrolled in courses they teach. In the dyslexic-children experiment, Williams and her colleagues did not randomly sample all dyslexic children but instead took an available group of subjects and randomly assigned them to the two different conditions.

random assignment any procedure that allows chance to divide an available group of subjects or observations into two or more subgroups

In the case of testing hypotheses about the difference between *two* means, the random assignment of available subjects is generally used as a substitute for random sampling. **Random assignment** is a method for dividing an available group of subjects or observations into two or more subgroups. It refers to any set of procedures that allows chance to determine who is included in what subgroup. For example, a researcher might assign available subjects to one of two groups by tossing a coin. The random-assignment model differs from the random-sampling model in that there is no well-defined population from which the samples are selected.

A problem arises because the procedures for inference for the two models are different. Those required for the random-assignment model are more burdensome and require the use of a computer to produce an exact solution. (We discuss these procedures in Section 21.2.) Now for the good news: *For most situations in which groups are formed by random assignment, application of the random-sampling model will lead to the same statistical conclusion as would result from application of the proper model* (Boik, 1987; McHugh, 1963). For this reason, you will find that in the research literature, the simpler random-sampling model is almost always applied, even when random assignment has been used to form treatment groups. In the latter situation, for example, the t test can be and is used to account for chance sampling variation in the sample results.

There is, however, a difference between the two models that is of great importance. *With random assignment of available subjects, statistical conclusions do not go beyond the exact subjects studied and the exact set of conditions at the time*. This does not mean that broader generalizations (research conclusions) cannot properly be made. It means only that *statistics* does not provide a sufficient basis for making them. Research conclusions must be made on the basis of knowledge and understanding of the research area.

14.13 Random Sampling and Random Assignment as Experimental Controls

We have grown well acquainted with the role of random sampling in making it possible to know the characteristics of sampling distributions. When two or more groups are compared, random sampling serves a second function not always fully appreciated: experimental control.

Consider a study in which an investigator wishes to investigate the accuracy of perception under two different conditions. The two samples will consist of the two sets of perception scores. How will the investigator obtain these two samples? If the subjects are assigned by casual methods to one group or the other, there may well be factors associated with the assignment that are related to the performance being tested. If subjects first to volunteer are placed in one group, the two groups may differ in level of motivation and so subsequently in performance. If subjects are inspected as they report for the experimental chore and assigned by whim, the experimenter may unconsciously tend to assign those without glasses to the condition where better performance is expected. In these examples, motivation and visual acuity would be considered confounding variables. Any systematic differences among participants across conditions could serve as a confounding variable, making any causal interpretations ambiguous (if not impossible). For example, differences in accuracy of perception could be due to the two conditions, due to differences in motivation levels, or due to differences in visual acuity. We cannot identify which is the correct causal interpretation.

The primary experimental (as opposed to statistical) benefit of random sampling lies in the chance (and therefore impartial) assignment of extraneous influence between the groups to be compared. Those who are likely to do well have just as much chance of being in one treatment group as in another, and the same is true of those who are likely to do poorly. The beauty of random sampling is that it affords this type of experimental control over extraneous influences whether or not they are known to exist by the experimenter. Similarly, random assignment of available subjects provides experimental control over extraneous factors that might bias the results. It provides an impartial manner of assigning extraneous influences between the groups to be compared. How important is this? In a review of outcome studies of marital and family therapy, Shadish and Ragsdale (1996) found that studies that used random assignment of subjects resulted in consistently higher and less variable posttest effects than studies that used nonrandom assignment. Moreover, as we noted in Section 8.10, inferring valid causal relationships among variables is based on experimental design issues. When designing experiments, random assignment of participants to different conditions prior to introducing a manipulation (e.g., intervention, treatment, etc.) provides a stronger basis for causal inferences than the use of nonrandom assignment (Shadish, Cook, & Campbell, 2002; Stone-Romero & Rosopa, 2008).

Many independent variables of potential interest are not subject to manipulation by experimenters. In some cases, it is because we are unwilling to do so. For example, the effect of varying brain damage among human subjects must be studied by examining individuals to whom damage has occurred through injury or disease. Other variables are unmanipulable because they are inherent characteristics of the organism. Among such variables are ethnic origin, gender, and age. If we wish to study the effect of differences in such a variable, we must identify subpopulations possessing the desired differences. For example, if we wish to study the effect of differences in age on a particular learning task, we might obtain the correlation between age and task performance. Alternatively, we might identify subpopulations described as young, middle-aged, and elderly, draw samples from each, and compare their mean performance on the learning task.

In studies like these, a significant degree of control is lost. This loss of control makes it more difficult to interpret the outcome of such studies. For example, suppose we learn that only 50% of chemistry majors marry by age 30, whereas 80% of music majors do so. Statistical procedures that assume random sampling, such as t, can often provide useful information in circumstances like this if we are careful in our interpretations. Although our t test might indicate that the difference is not due to chance, it is completely unjustified to conclude that marital practice is owed to differential educational experience. What about the factors of personality, interest, and intellect that led these individuals to these educational paths? Had it been possible to assign subjects at random to the two majors, interpretation of the outcome of the study would be clearer. However, some of the most important empirical questions are not approachable in any other way.

14.14 Summary

Many questions asked in the behavioral sciences require measuring a dependent variable under two different conditions. The techniques that you learned about in this chapter apply only to samples that are obtained independently and at random. Samples are drawn from two populations and the difference between the sample means, $(\overline{X} - \overline{Y})$, is determined. If the experimental manipulation of the variable under study has no effect, then there should be no difference between the two population means; that is, $\mu_X - \mu_Y = 0$. However, even when this is true, we do not expect $\overline{X} = \mu_X$ or $\overline{Y} = \mu_Y$ because of random sampling variation, and thus we do not expect $\overline{X} - \overline{Y} = \mu_X - \mu_Y$. But if H_0: $\mu_X - \mu_Y = 0$ is true, the mean of an infinite number of $(\overline{X} - \overline{Y})$s will be equal to zero (the value stated in the null hypothesis) and so we expect our value of $(\overline{X} - \overline{Y})$ to be close to zero.

If the sampling distributions of $\overline{X}$ and $\overline{Y}$ are normal, the distribution of an infinite number of $(\overline{X} - \overline{Y})$s selected at random will also be normal. As predicted by the central limit theorem, the distribution of $(\overline{X} - \overline{Y})$ approximates a normal curve as sample size increases even when X and Y are not normally distributed. We could convert an actual $(\overline{X} - \overline{Y})$ obtained from sampling to a z value and use the normal curve table if we knew the standard error of the difference, $\sigma_{\overline{X}-\overline{Y}}$, but this requires that we know σ_X and σ_Y. As in tests of a single mean, we must estimate the two population standard deviations from the sample data. The result is $s_{\overline{X}-\overline{Y}}$, the estimated standard error of the difference between sample means. To test differences between means, we assume homogeneity of variance and obtain a pooled estimate of $\sigma_{\overline{X}-\overline{Y}}$ with $(n_X - 1) + (n_Y - 1)$ degrees of freedom. Protection from unfortunate consequences of violating the assumption of homogeneity of variance is best assured by using large samples of equal size.

We obtain a t value by dividing the deviation of $(\overline{X} - \overline{Y})$ from $(\mu_X - \mu_Y)_{\text{hyp}}$ by the estimate of $\sigma_{\overline{X}-\overline{Y}}$:

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{s_{\overline{X}-\overline{Y}}} = \frac{\begin{matrix}\textit{obtained difference}\\ \textit{between sample}\\ \textit{means}\end{matrix} - \begin{matrix}\textit{hypothesized difference}\\ \textit{between population means}\end{matrix}}{\begin{matrix}\textit{estimated standard error of the difference}\\ \textit{between two means}\end{matrix}}$$

If the resultant t value falls in the region of rejection (which is determined by the nature of H_A, the chosen value of α, and the particular distribution of t related to df), we reject H_0 and conclude that the two samples were not drawn from two populations having the same mean.

One of the basic assumptions that allow us to make inferences about population parameters from sample statistics is that of random sampling. In comparisons of two (or more) conditions, the random assignment of available subjects is often used as a substitute for random sampling. Although the t test can be used in this situation, statistics does not provide the basis for making generalizations beyond the exact subjects studied and the exact set of conditions at the time.

Key Terms, Concepts, and Symbols

independent samples (p. 233)
random sampling distribution of $(\overline{X} - \overline{Y})$ (p. 234)
$\mu_{\overline{X}-\overline{Y}}$ (p. 234)
standard error of $(\overline{X} - \overline{Y})$ (p. 236)
$\sigma_{\overline{X}-\overline{Y}}$ (p. 236)
estimated standard error of $(\overline{X} - \overline{Y})$ (p. 237)
$s_{\overline{X}-\overline{Y}}$ (p. 237)
assumption of homogeneity of variance (p. 237)
pooled estimate of the population variance s_p^2 (p. 238)
degrees of freedom (p. 238)
formula for t, unequal n (p. 239)
formula for t, equal n (p. 239)
effect size (p. 246)
d (p. 246)
g (p. 246)
r (p. 248)
power (p. 249)
central limit theorem (p. 251)
random-sampling model (p. 251)
random assignment (p. 252)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Forty subjects are randomly assigned to receive either a new memory pill or a placebo. After treatment, they are all given a difficult recall test. The mean for the memory-pill group is 82%, whereas that for the placebo group is 64%. Can we conclude that the memory pill worked? Explain.

2. Given: independent samples with $s_X = s_Y = 10$. (a) Calculate $s_{\overline{X}-\overline{Y}}$ from Formula 14.2*b*, assuming $n_X = n_Y = 30$. (b) Recalculate $s_{\overline{X}-\overline{Y}}$, assuming $n_X = 15$ and $n_Y = 45$. (c) Compare the two values of $s_{\overline{X}-\overline{Y}}$. What principle does this outcome illustrate?

3. Given: $\overline{X} = 165.0$, $SS_X = 840$, $n_X = 36$, $\overline{Y} = 175.0$, $SS_Y = 1008$, and $n_Y = 49$; assume the samples are independent. (a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Calculate $s_{\overline{X}-\overline{Y}}$. (c) Calculate t. (d) Evaluate t for $\alpha = .05$ and $\alpha = .01$, and state your conclusions.

4. You are given the following scores from two randomly drawn independent samples:

X: 6, 7, 8, 8, 11 *Y*: 3, 4, 4, 7, 7

(a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Complete the test at the .05 and .01 levels of significance, and state your conclusions.

5. Given: $\mu_X = 100$, $\sigma_X = 16$, $\mu_Y = 104$, $\sigma_Y = 16$. Compute d and make a rough sketch of the populations to show the degree of separation (or overlap).

6. Professor Jones conducts a study to compare two different dosages of AZT and obtains $\sigma_1 = \sigma_2 = 5.2$ and $d = .68$ for the response measure. How far apart are μ_1 and μ_2?

7. Many studies have found a negative correlation between amount of exercise and level of depression. People who exercise tend to be less depressed than people who do not exercise. Recall from Chapter 7, however, that correlation does not *prove* causation. (Does exercising regularly lower depression, or are less depressed individuals more likely to exercise?) Suppose we wished to investigate the effects of exercise on depression. We randomly assign 20 subjects, who tests show suffer from depression, to one of two groups of 10: those who will participate in an aerobic exercise program 4 days a week for 2 months, and those who are not prescribed an exercise program. A study similar to this was, in fact, conducted by McCann and Holmes (1984).

Let us assume that one of the subjects in the exercise group and two of those in the control group drop out before completing the study. At the end of the study, the hypothetical depression scores for the remaining subjects are as follows (higher scores indicate higher levels of depression):

EXERCISE	NO EXERCISE
19	25
8	32
25	27
17	35
15	38
17	26
27	35
13	42
21	

(a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Complete the test at the .05 level of significance, and state your conclusion. (c) Does it matter which group we designate as *X* and which as *Y*? (d) Compute the estimated effect size, *g*, and *r*.

8. (a) Did the experiment described in Problem 7 employ random sampling or random assignment? (b) What technique of introducing randomness into research does the *t* test assume? (c) How do you resolve the discrepancy?

9. At the start of Chapter 1 we cited the study by Fischl et al. (1987) about the effects of the drug AZT in patients with AIDS. Suppose you were assigned to investigate the effects of a new experimental AIDS drug called ABZ on CD4+ cell counts in blood (the virus that causes AIDS kills the CD4+ cells). You randomly assign 20 AIDS patients who have CD4+ cell counts of fewer than 200 to one of two groups: those receiving ABZ for 90 days and those receiving a placebo (an inert substance) for the same period. Imagine that the CD4+ cell counts at the end of 90 days are as follows:

ABZ GROUP	PLACEBO GROUP
240	160
300	195
195	180
230	150
250	170
310	175
270	180
290	165
180	175
210	190

(a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) The 10 individuals given ABZ are the only 10 people ever to have received ABZ; are they the entire population or a sample? (c) Complete the test at the .01 level of significance, and state your conclusion. (d) Compute the estimated effect size, *g*, and *r*. (e) Is it really possible to randomly select 20 AIDS patients from all those individuals who have AIDS? How might this affect the conclusions?

10. Albert Bandura has conducted a number of studies on aggression in children. In one study (Bandura, Ross, & Ross, 1963), one group of nursery school children were shown a film of adults punching, hitting (with a hammer), kicking, and throwing a large rubber Bobo doll. Another group was not shown the film. Afterward, both groups were allowed to play with Bobo dolls in a playroom, and the number of violent contacts were measured in each child. Let us imagine that we replicated this study and obtained the following data (scores are the number of violent contacts with the Bobo doll):

FILM GROUP	NO FILM GROUP
20	5
65	20
41	0
80	0
52	10
35	8
15	30
75	13
60	0
50	25
47	
33	

(a) State formally the hypotheses necessary to conduct a directional test of no difference between the two population means. (b) Complete the test at the .01 level of significance, and state your conclusion. (c) Compute the estimated effect size, *g*, and *r*.

11. We wish to know whether supplementing the diet with extra quantities of vitamin A makes a difference in ability to see under conditions of dim illumination. We randomly assign 20 subjects to one of two groups, and feed one group a normal diet and the other a normal diet plus supplementary vitamin A. After 1 week under this regimen, each individual is tested for visual acuity in conditions of dim illumination. The following visual acuity scores are obtained:

SUPPLEMENTARY VITAMIN A (X)	NORMAL DIET (Y)
38	37
41	34
42	45
47	40
42	43
45	40
48	42
31	22
38	28
30	31

(a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Complete the test at the .05 level of significance, and state your conclusions. (c) Does the result mean that H_0 is true? Explain. (d) Compute the estimated effect size, *g*, and *r*.

12. A psychologist wishes to compare mean reaction time on a frequently used reaction task with that on a new task. He assumes that the standard deviation will be 15 msec on both tasks and wants to estimate at the 95% level of confidence the difference between the two population means. The psychologist wants this difference to be within 5 msec of the difference between the sample means. If the psychologist selects two random samples, what should be the size of each?

13. A psychologist wonders whether midsemester warning notices affect performance. She decides to select a sample of delinquent students and, at random, to send such notices to half of them and no notices to the other half. Experience suggests that among such delinquent students, $\sigma_{\text{GPA}} = .30$. She decides to adopt $\alpha = .05$ and power $= .95$. If the difference between the warned and unwarned students is as great as .075 grade points, she would want to know it. (a) State the null hypothesis and the alternative hypothesis for a two-tailed test. (b) What should the size of the sample be to test this null hypothesis? Use Cohen's power curves.

14. In reference to Problem 13, what should the sample size be if the psychologist feels that: (a) She should choose power = .80? (b) Is it important to know whether the difference between the means is .15 grade point (or more) and power = .95?

15. Given that $\mu_X = 30$, $\sigma_X = 4$, $\mu_Y = 33$, and $\sigma_Y = 4$, if you conduct a two-tailed test of H_0: $\mu_X - \mu_Y = 0$ with $\alpha = .05$, how many times in 50 replications would you expect to obtain significant results for samples of size (a) 10? (b) 20? (c) 50? (d) 100?

16. For a two-tailed test of H_0: $\mu_X - \mu_Y = 0$ with $\alpha = .05$ and $n = 10$, what is the probability of detecting a d of (a) .2? (b) .5? (c) .8? (d) What does this tell you about using samples of this size?

17. For a two-tailed test of H_0: $\mu_X - \mu_Y = 0$ with $\alpha = .05$ and $n = 100$, what is the probability of detecting a d of (a) .2? (b) .5? (c) .8? (d) What does this tell you about using samples of this size? (Compare with Problem 16.)

18. Professor Smith wishes to compare the performance of students in her class on multiple-choice and fill-in-the-blank exams. She randomly chooses half the class for each test. If each test has a standard deviation of 10, what is the probability of detecting a true difference ($\mu_X - \mu_Y$) of 5 points with a two-tailed test, $\alpha = .05$, when the sample size is (a) 10? (b) 25? (c) 50? (d) 100?

19. For Problem 18, if Professor Smith wishes power = .8, what size sample would she need to detect a true difference of (a) 2 points? (b) 5 points? (c) 8 points? (d) 12 points?

20. Self-fulfilling prophecies refer to beliefs or expectations that can bring about enhanced (or impaired) performance. Research by Eden (2003) and his colleagues have demonstrated how self-fulfilling prophecies can affect actual job performance. For example, Eden and Zuk (1995) randomly assigned naval cadets to a control and experimental condition, with those in the experimental condition led to believe that they could overcome seasickness and perform well even during rough seas. Their results showed that those in the experimental condition actually reported less seasickness and performed better than cadets who were assigned to the control condition. Suppose you replicated their study. Below are performance scores from the experimental (X) and control (Y) conditions. (a) Use IBM SPSS to conduct a two-tailed hypothesis test with $\alpha = .01$. (b) Although IBM SPSS does not report effect size r, use information in the IBM SPSS output to calculate r.

Experimental (X)		**Control (Y)**	
90	86	77	80
91	91	82	79
92	92	79	81
90	90	80	82
91	91	79	79
92	91	77	82
91	91	78	86
90	84	78	78
90	90	80	80
90	89	83	80
88	89	82	81
90	85	79	82
87	84	80	79
			81

CHAPTER 15

Testing for a Difference between Two Dependent (Correlated) Groups

When you have finished studying this chapter, you should be able to:

- Describe the different ways in which dependent samples may be generated;
- Understand the nature of the random sampling distribution of $(\overline{X} - \overline{Y})$ and $\overline{D}$ for dependent samples;
- Calculate $s_{\overline{X}-\overline{Y}}$ for dependent samples and explain how it differs from the estimated standard error of the difference for independent samples;
- Calculate t by the computational formula and with IBM SPSS;
- Calculate effect size for the difference between two dependent means;
- Understand the underlying assumptions necessary for testing hypotheses about the difference between two dependent means; and
- Recognize some potential problems with using the dependent-groups design.

In Chapter 14, we considered differences obtained from two independent random samples, where measurements in the two samples are unrelated to each other. It is also possible to conduct a study in which measurements are related. This is called a **dependent-samples design**.

dependent-samples design
a study in which measurements in one sample are related to measurements in the other sample(s)

In what circumstances would samples be dependent? Consider our hypothetical vitamin A experiment in Problem 11 of the previous chapter. In that problem, we wished to know whether supplementing the diet with extra quantities of vitamin A makes a difference in the ability to see under conditions of dim illumination. Instead of selecting two groups of 10 subjects at random and feeding one a normal diet and the other a diet supplemented with vitamin A, we could have selected one group of 10 subjects at random, fed half of them a normal diet and fed the other half a normal diet supplemented with vitamin A, tested their vision, and then retested them under the condition of the other diet. This is a **repeated-measures design**. In this case, we would have 10 *pairs* of observations, with each pair measured on the *same* subject. The potential advantage of this design is that it can reduce differences between the two groups of scores due to random sampling. Each subject serves as his or her own control. Presumably, a subject who was relatively good under one condition would tend to be relatively good under the other, aside from possible difference due to vitamin A. Thus, for a given pair of observations, the value of Y is in part determined by (or related to) the particular value of X, and the samples cannot be said to be independent.

repeated-measures design
a study in which observations are measured on the same subject under two (or more) conditions

Samples may be dependent even when different subjects have been used. In designing the vitamin A experiment, for example, we probably would consider that an extraneous source of variation is contributed by differences in basic visual acuity among individuals selected at random. We could eliminate this source of variation by pairing individuals matched on pretest basic visual acuity. This is called a **matched-subjects design**. Thus, the two members of the first pair of subjects may each have 20–40 vision, the members of the second pair may each have 20–20 vision, and so forth. When observations are paired in this way, the value of any particular Y score will be in part related to the value of its paired X score, and so the values of X and Y cannot be said to be completely independent.

matched-subjects design
a study in which subjects from two (or more) samples are paired (matched) by the experimenter on an extraneous variable

In a matched-subjects design, it is not always necessary that the experimenter match subjects on some variable to form pairs. Some pairs of observations occur naturally in the population. For example, investigators may choose to study identical twins, fathers and sons, mothers and daughters, or husbands and wives. Many people prefer to call such studies **matched-pairs investigations** (rather than "matched-pairs experiments").

matched-pairs investigations
a study in which subjects from two groups are naturally paired

There are, then, three basic ways in which dependent samples may be generated: (1) the same subjects are used for both conditions of the study; (2) different subjects are used, but they are matched by the experimenter on some variable related to performance on the variable being observed; and (3) natural pairings.

15.1 Determining a Formula for *t*

For dependent means, the ideal model supposes that there exists a population of paired observations (through matching or repeated measures) and that a sample is selected at random from this population. It also presupposes that we know ρ (rho), the correlation between pairs of observations in the population. The standard error of the difference between dependent means takes account of the correlation induced by the existing relationship between the samples. It is:

STANDARD ERROR OF THE DIFFERENCE
BETWEEN TWO DEPENDENT MEANS

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\sigma^2_{\overline{X}} + \sigma^2_{\overline{Y}} - 2\rho_{XY}\sigma_{\overline{X}}\sigma_{\overline{Y}}} \qquad (15.1)$$

where: $\sigma_{\overline{X}}$ = standard error of the mean of X
$\sigma_{\overline{Y}}$ = standard error of the mean of Y
ρ_{XY} = correlation coefficient for the population of pairs of X and Y measures.

$\boldsymbol{s_{\overline{X}-\overline{Y}}}$
symbol for the estimated standard error of $(\overline{X} - \overline{Y})$

Notice that Formula 15.1 is identical to Formula 14.1 (the standard-error formula for independent samples) except for the last term, $(-2\rho_{XY}\sigma_{\overline{X}}\sigma_{\overline{Y}})$. If $\rho = 0$, the last term drops out completely. When the parameters are unknown, as usually is the case, the formula that estimates $\sigma_{\overline{X}-\overline{Y}}$ is:

$\boldsymbol{\sqrt{s^2_{\overline{X}} + s^2_{\overline{Y}} - 2rs_{\overline{X}}s_{\overline{Y}}}}$
formula for the estimated standard error of $(\overline{X} - \overline{Y})$ in the dependent-samples design

ESTIMATED STANDARD ERROR
OF THE DIFFERENCE BETWEEN
TWO DEPENDENT MEANS

$$s_{\overline{X}-\overline{Y}} = \sqrt{s^2_{\overline{X}} + s^2_{\overline{Y}} - 2rs_{\overline{X}}\,s_{\overline{Y}}} \qquad (15.2)$$

As with independent samples, the **estimated standard error for dependent samples** varies from sample to sample. Therefore, we must convert the difference between two means to a t value rather than a z value. The calculation of t for a test of the difference between two dependent means is:

$\boldsymbol{t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{s_{\overline{X}-\overline{Y}}}}$
formula for the t statistic for testing a null hypothesis about the difference between two population means

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{\sqrt{s^2_{\overline{X}} + s^2_{\overline{Y}} - 2rs_{\overline{X}}\,s_{\overline{Y}}}} \qquad (15.3)$$

What is the advantage of Formula 15.3 over the t test for independent samples? Look at the denominator in Formula 15.3. You can see that the higher the value of r (the correlation between the paired scores), the smaller the value of $s_{\overline{X}-\overline{Y}}$ becomes. This will increase the value of t (assuming the numerator does not change) and gives us a greater chance of rejecting a false null hypothesis.

15.2 Degrees of Freedom for Tests of No Difference between Dependent Means

$\boldsymbol{n - 1}$
degrees of freedom for t in the dependent-samples design, where n equals the number of pairs of scores

After we have calculated t, we evaluate t with

$$df = n - 1$$

where n is the number of *pairs* of scores. Note that $df = (n - 1)$ is half of the $(n_X - 1) + (n_Y - 1)$ degrees of freedom that we would use if the same study were conducted using independent samples. Why are there fewer degrees of freedom for dependent samples?

When samples are independent, the score recorded for the first subject in the first group and that recorded for the first subject in the second group are completely unrelated. But

D
symbol for a difference score, $X - Y$

$\overline{D}$
symbol for the mean of a sample of difference scores

when the same subjects are used under both treatment conditions, or when subjects are matched, this is not so. In the repeated-measures design, subjects who do well under one condition tend to do well under the second condition, and those who do poorly under one condition tend to do poorly under the other also. In the matched-subjects design, if a subject does very well as compared with the others under one condition, his or her matched counterpart will tend to do well as compared with the others under the other condition if the matching has been on a relevant variable. In short, performance of one member of the pair is *not* independent of that of the other member. Thus, *when the score of one member of the pair is specified, that of the other is not completely free to vary.* Consequently, only one degree of freedom can be ascribed to each pair of scores. Because there are n pairs of scores, there will be $n - 1$ degrees of freedom.

15.3 An Alternative Approach to the Problem of Two Dependent Means

sampling distribution of $\overline{D}$
a theoretical relative frequency distribution of all the possible values of $\overline{D}$ that would be obtained by chance from an infinite number of samples of a particular size drawn from two populations (X and Y)

s_D
symbol for the estimated population standard deviation of D

Formula 15.3 requires that we compute the correlation between the paired values of X and Y. This is very time consuming if you are making calculations by hand. There is an alternative method for calculating t for the test of two dependent means that provides an identical outcome with less computation. This method focuses on the characteristics of the **sampling distribution of differences** between the paired X and Y scores.

Consider the null hypothesis that $\mu_X - \mu_Y = 0$. If the hypothesis is true, then it is also true that the mean of the population of differences between the paired values of X and Y is zero. If the difference between paired X and Y scores is designated by **D**, the initial hypothesis may be restated as: H_0: $\mu_D = 0$. With this method, we find **$\overline{D}$**, the mean of the sample set of different scores, and inquire whether it differs significantly from the hypothesized mean of the population of difference scores, μ_D. Notice that *this transforms the test from a two-sample test (of* X *and* Y *scores) to a one-sample test (of difference scores).*

Instead of a sample of scores, X, we have a sample of difference scores, D. The **estimated standard deviation of the difference scores**, **s_D**, is:

$$s_D = \sqrt{\frac{SS_D}{n-1}} = \sqrt{\frac{\sum(D-\overline{D})^2}{n-1}} = \sqrt{\frac{\sum D^2 - \frac{\left(\sum D\right)^2}{n}}{n-1}}$$

$$\sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}}$$

formula for the estimated population standard deviation of D

The **estimated standard error of the mean of the difference scores, $s_{\overline{D}}$**, is thus:

$$s_{\overline{D}} = \frac{s_D}{\sqrt{n}}$$

and our formula for t becomes:

FORMULA FOR t FOR TESTING HYPOTHESES
ABOUT DEPENDENT MEANS

$$t = \frac{\overline{D} - \mu_{D(\text{hyp})}}{s_{\overline{D}}} \qquad (15.4)$$

$s_{\overline{D}}$ symbol for the estimated standard error of $\overline{D}$

$t = \dfrac{\overline{D} - \mu_{D(\text{hyp})}}{s_{\overline{D}}}$ alternative formula for the t statistic for testing a null hypothesis about the difference between two dependent means

A computational formula for t for dependent means is easily derived from the preceding formulas:

COMPUTATIONAL FORMULA FOR t FOR DEPENDENT MEANS

$$t = \frac{\overline{D} - \mu_{D(\text{hyp})}}{\sqrt{\dfrac{\sum D^2 - \dfrac{\left(\sum D\right)^2}{n}}{n(n-1)}}} \tag{15.5}$$

Compare Formula 15.5 with Formula 12.5 in Section 12.14, the computational formula for t for testing hypotheses about single means.

15.4 Testing a Hypothesis about Two Dependent Means: Does Text Messaging Impair Driving?

Human Factors Psychology applies basic research in order to develop technological systems that are safe, productive, comfortable, and error-free. This includes driving safety. Have you ever text messaged anyone while driving? Used a cell phone or iPod? Driver distractions are responsible for many automobile accidents (Eby & Kostyniuk, 2003). To investigate to what extent use of wireless technological devices may distract drivers, psychologists Matthew Crisler and Johnell Brooks and colleagues (2008) tested 14 licensed drivers in a fixed-base driving simulator in which the drivers controlled the steering wheel and pedals. The simulated road on which they drove was a very curvy two-lane rural road. Each lane was 9.5 feet wide. The 14 participants had to drive 55 mph under two conditions: (1) with no distracting task, and (2) while text messaging (the experimenters answered each message and responded with a new question). The dependent variable was percentage of time driving within the boundaries of the lane.

Crisler, Brooks, and colleagues used a repeated-measures design: half the subjects were first tested under the no distraction condition and then were later tested under the text messaging condition, and vice-versa for the other subjects. *Which subject was assigned to which condition first was determined at random.*

Crisler et al. tested H_0: $\mu_X - \mu_Y = 0$ (where X represents the percent of time driving within the lane during no distractions and Y the percent of time driving within the lane while text messaging against H_A: $\mu_X - \mu_Y \neq 0$, using the 5% significance level. When restated in terms of difference scores, the null hypothesis becomes H_0: $\mu_D = 0$.

We illustrate the use of Formula 15.5 in Table 15.1 (the data are real). To conduct the t test, you subtract each value of Y from its paired value of X and record the difference in a new column, D (Step ①). The only quantities needed for the calculation of t (Step ⑤) are ΣD, ΣD^2, and $\overline{D}$ (Steps ②, ③, and ④, respectively).

The calculated value of t is 4.12. The sampling distribution of t when H_0 is true is shown in Figure 15.1. The expected value is 0; the critical value of t for a two-tailed test with $\alpha = .05$ and $df = 13$ is ± 2.160 (from Table D in Appendix D). The probability of drawing samples this different from the value stated in H_0 is less than 1 in 20 when H_0 is true. The obtained t falls in the region of rejection, and therefore Crisler et al. rejected H_0.

TABLE 15.1 *Test of the Difference between Two Dependent Means by Formula 15.5*

	PERCENT TIME IN LANE			
PAIR	NO DISTRACTION X	WHILE TEXT MESSAGING Y	① $D = X - Y$	D^2
1	94.62	61.63	32.99	1088.34
2	82.89	82.03	.86	.74
3	96.78	75.42	21.36	456.25
4	96.85	85.08	11.77	138.53
5	96.28	77.34	18.94	358.72
6	99.34	94.19	5.15	26.52
7	95.88	93.15	2.73	7.45
8	86.05	89.74	−3.69	13.62
9	94.85	91.10	3.75	14.06
10	99.30	91.26	8.04	64.64
11	98.81	87.68	11.13	123.88
12	99.27	92.43	6.84	46.79
13	93.32	70.93	22.39	501.31
14	99.27	88.37	10.90	118.81
			$\sum D = 153.16$ ②	$\sum D^2 = 2959.66$ ③

Calculations:

④ $$\bar{D} = \frac{\sum D}{n} = \frac{153.16}{14} = 10.94$$

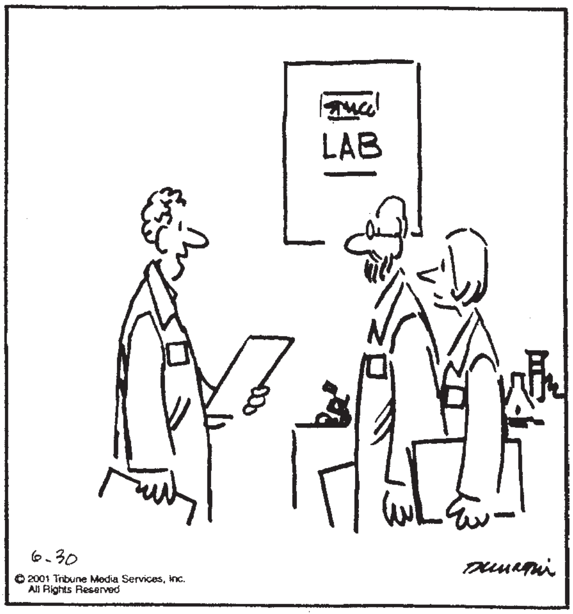
⑤ $$t = \frac{\bar{D} - \mu_{D(\text{hyp})}}{\sqrt{\dfrac{\sum D^2 - \dfrac{\left(\sum D\right)^2}{n}}{n(n-1)}}}$$

$$= \frac{10.94 - 0}{\sqrt{\dfrac{2959.66 - \dfrac{23{,}457.98}{14}}{14(13)}}}$$

$$= \frac{10.94}{\sqrt{7.055}}$$

$$= 4.12$$

Dunagin's People



"The placebo proved more effective than our product ... let's market the placebo!"

© Tribune Media Services, Inc. All rights reserved.
Reprinted with permission.

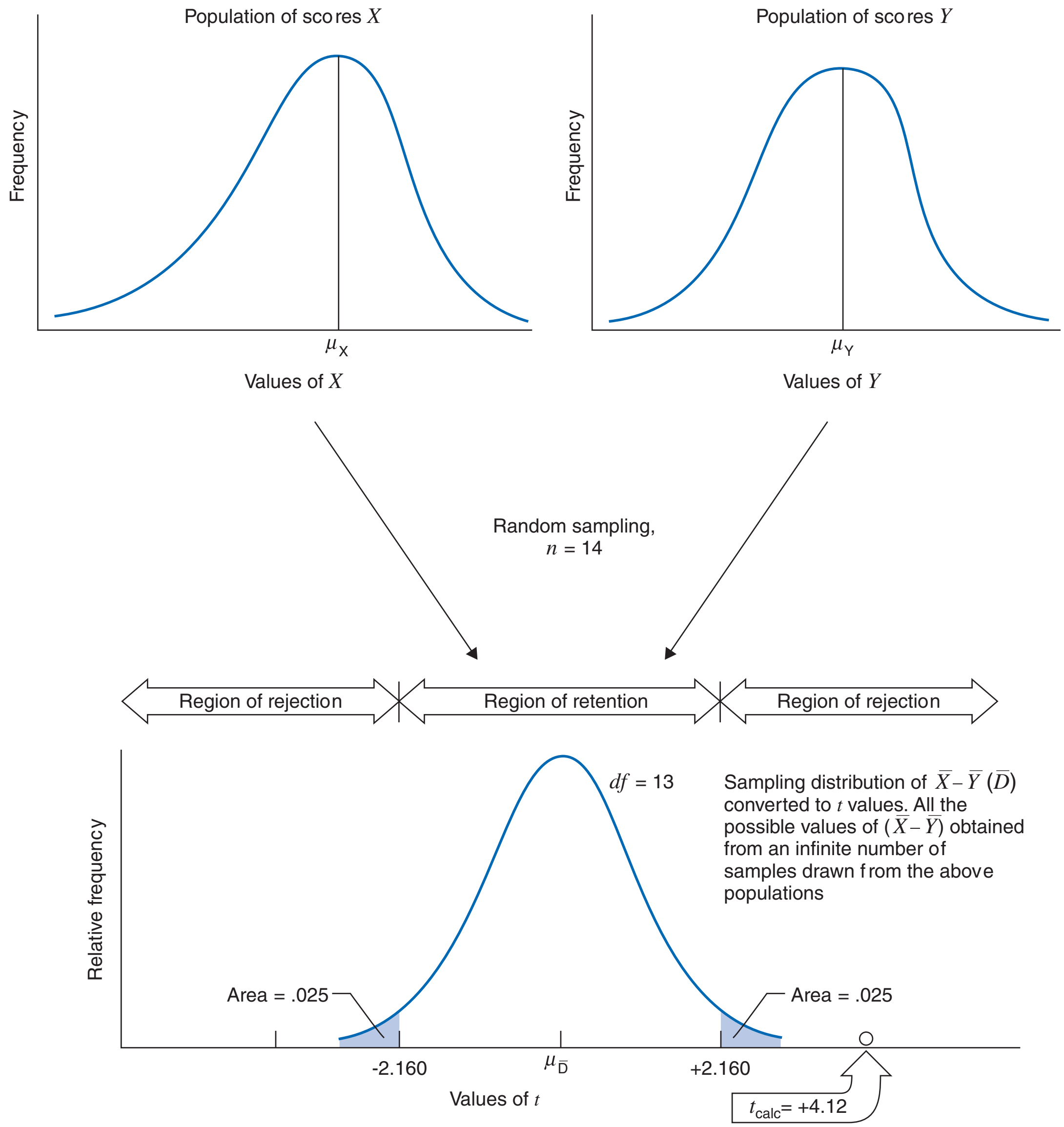

Population of scores X
Frequency
μ_X
Values of X
Population of scores Y
Frequency
μ_Y
Values of Y
Random sampling,
n = 14
Region of rejection
Region of retention
Region of rejection
Relative frequency
df = 13
Sampling distribution of $\overline{X} - \overline{Y}$ ($\overline{D}$) converted to t values. All the possible values of ($\overline{X} - \overline{Y}$) obtained from an infinite number of samples drawn from the above populations
Area = .025
Area = .025
-2.160
$\mu_{\overline{D}}$
+2.160
Values of t
t_{calc}= +4.12

FIGURE 15.1 Testing a hypothesis about the difference between two dependent means against a nondirectional alternative; $\alpha = .05$.

This was the statistical conclusion. The *research conclusion* is that people's driving ability is worse when they are text messaging.

Although Formula 15.5 reduces the computational burden substantially, it also yields less information than Formula 15.3. When we are done, we know the size of the difference between the two sample means ($\overline{D}$) and whether the difference was or was not significant according to the decision criterion adopted. In most research, we will want to know (and will be obliged to report) the two means and the two standard deviations. If we are curious about how much correlation was induced by pairing, we will also want to know r. The short method yields none of this information. If these quantities are desired, we will have to return to the data and compute them.

15.5 Calculating *t* with IBM SPSS

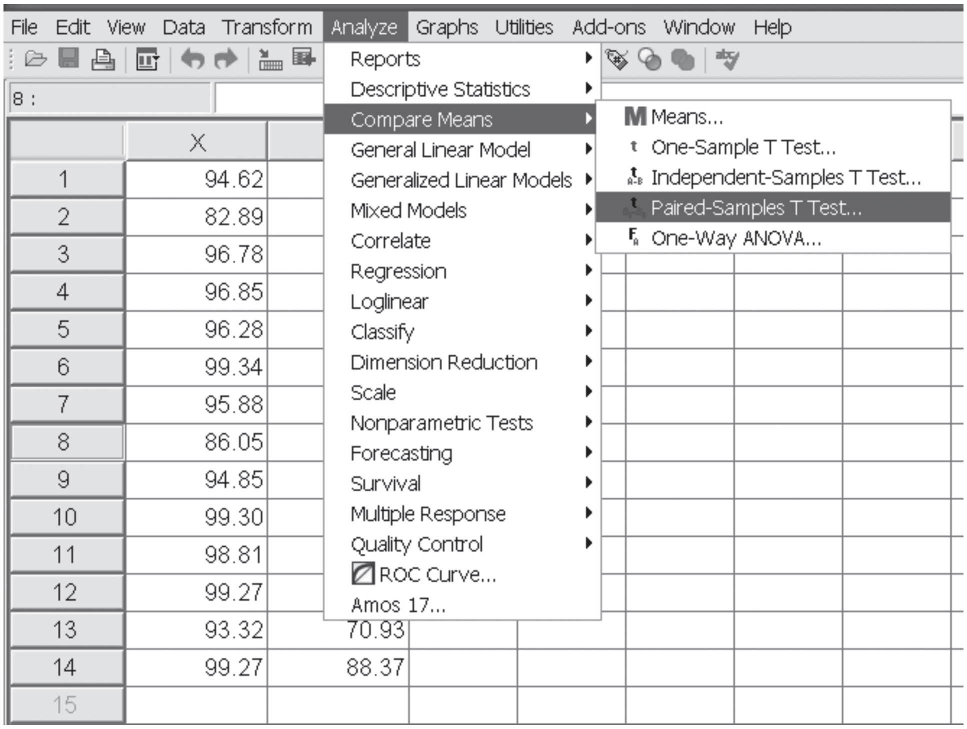
In this section, you will learn how to use IBM SPSS to calculate the *t* statistic to test hypotheses about two dependent means. You will use the data in Table 15.1.

Recall that with dependent groups there are *n* pairs of scores. In the **Data Editor**, you must enter the data in two columns—one for each group. So, when entering the data in IBM SPSS, there will be 14 rows of scores in two columns. In the **Variable View**, for the first variable, we will use *X* as the variable name and Percent Time in Lane with No Distraction as the label. For the second variable, we will use *Y* as the variable name and Percent Time in Lane while Text Messaging as the label. We believe that it is important to provide detailed labels especially in instances when you have to access your data weeks or months later. The labels can help in clarifying what each variable is.

To conduct a hypothesis test on two dependent means in IBM SPSS, click **Analyze** then select **Compare Means** followed by **Paired-Samples T Test**. Figure 15.2 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze → Compare Means → Paired-Samples T Test**.

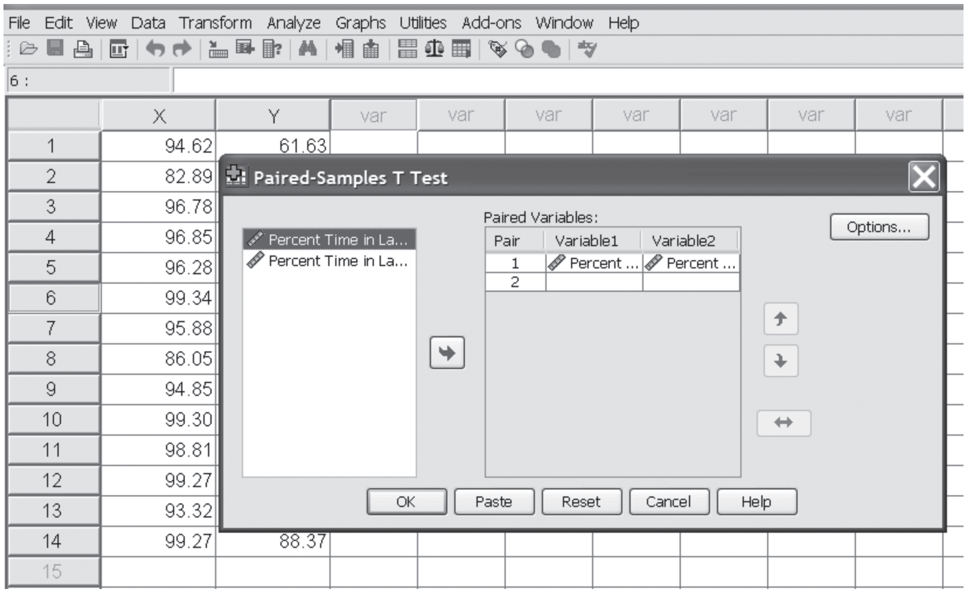
Now, a dialog box will appear. You should move the Percent Time in Lane with No Distraction variable (i.e., *X*) to Variable 1 in the Paired Variables box. Next, move the Percent Time in Lane while Text Messaging variable (i.e., *Y*) to Variable 2 in the Paired Variables box. Figure 15.3 shows a screenshot after the variables have been moved. Now, click OK.

Table 15.2 shows the IBM SPSS output. After the syntax, there is a section of output labeled Paired Samples Statistics. As you can see, it contains various statistics, but for each group. For example, we see that 95.25 is the mean percent time in lane when driving with no distraction compared to 84.31 when driving while text messaging. IBM SPSS also provides the estimated standard deviation and estimated standard error of the mean in each group. The next


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
8 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...

	X	
1	94.62	
2	82.89	
3	96.78	
4	96.85	
5	96.28	
6	99.34	
7	95.88	
8	86.05	
9	94.85	
10	99.30	
11	98.81	
12	99.27	
13	93.32	70.93
14	99.27	88.37
15		

FIGURE 15.2 IBM SPSS menus when calculating *t* for dependent means.


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
6 :
Options...
OK
Paste
Reset
Cancel
Help

	X	Y	var	var	var	var	var	var	var
1	94.62	61.63							
2	82.89								
3	96.78								
4	96.85								
5	96.28								
6	99.34								
7	95.88								
8	86.05								
9	94.85								
10	99.30								
11	98.81								
12	99.27								
13	93.32								
14	99.27	88.37							
15									

Paired-Samples T Test

Percent Time in La...
Percent Time in La...

Paired Variables:

Pair	Variable1	Variable2
1	Percent ...	Percent ...
2		

FIGURE 15.3 IBM SPSS dialog box for selecting variables used in calculating *t* for dependent means.

TABLE 15.2 *IBM SPSS Output from Dependent Samples t*

```
T-Test pairs = X with Y (Paired)
  /Criteria = CI (.9500)
  /Missing = Analysis.
```

PAIRED SAMPLES STATISTICS

		MEAN	*N*	STD. DEVIATION	STD. ERROR MEAN
Pair 1	Percent Time in Lane with No Distraction	95.2507	14	5.00903	1.33872
	Percent Time in Lane while Text Messaging	84.3107	14	9.68837	2.58932

PAIRED SAMPLES CORRELATIONS

		N	CORRELATION	SIG.
Pair 1	Percent Time in Lane with No Distraction & Percent Time in Lane while Text Messaging	14	.208	.476

PAIRED SAMPLES TEST

		PAIRED DIFFERENCES							
					95% CONFIDENCE INTERVAL OF THE DIFFERENCE				
		MEAN	STD. DEVIATION	STD. ERROR MEAN	LOWER	UPPER	*t*	*df*	SIG. (2-TAILED)
Pair 1	Percent Time in Lane with No Distraction – Percent Time in Lane while Text Messaging	10.94000	9.93864	2.65621	5.20160	16.67840	4.119	13	.001

section of output is labeled Paired Samples Correlations. Here, we see that the correlation between X and Y is +.208. The next column provides the p-value associated with a two-sided t test of the null hypothesis that $\rho = 0$. This will be covered in Section 16.2. In the next section of output labeled Paired Samples Test, the first column shows the order of the difference in the numerator of the t. Here, it was based on $X - Y$. Next, we see that the mean difference was $(94.25 - 84.31) = 10.94$. The column labeled Std. Deviation refers to the estimated standard deviation of the difference scores (s_D) which is 9.93864. The column labeled Std. Error Mean actually refers to the estimated standard error of the mean of the difference scores ($s_{\overline{D}}$) which, in this case, is 2.65621. The next two columns will be discussed later in Chapter 17. The last three columns display, respectively, the calculated t, the df, and the two-sided p-value.

15.6 Effect Size

Look again at Section 14.9. Recall that for tests of hypotheses about the difference between two means, effect size d is

$$d = \frac{(\mu_X - \mu_Y)_{\text{true}} - (\mu_X - \mu_Y)_{\text{hyp}}}{\sigma}$$

where σ is the population standard deviation, not the standard error ($\sigma_{\overline{X}-\overline{Y}}$). Because we do not know σ, we must estimate it with s. Remember also that ordinarily $(\mu_X - \mu_Y)_{\text{hyp}} = 0$. Therefore,

$$\text{estimated } d = g = \frac{\overline{X} - \overline{Y}}{s}$$

For dependent samples, s, the estimated population standard deviation, is $s_{X-Y} = \sqrt{s_X^2 + s_Y^2 - 2r_{XY}\, s_X\, s_Y}$. Therefore,

(Hedges') *g* a measure of effect size; expresses the difference between $\overline{X}$ and $\overline{Y}$ relative to s_{X-Y} (or s_D), the estimated population standard deviations

$$\text{estimated } d = g = \frac{\overline{X} - \overline{Y}}{\sqrt{s_X^2 + s_Y^2 - 2r_{XY}\, s_X\, s_Y}} \tag{15.6}$$

As usual, it is much easier to estimate d if we use difference scores, D (remember, $\overline{X} - \overline{Y} = \overline{D}$):

$$\text{estimated } d = g = \frac{\overline{D}}{\sqrt{\dfrac{\sum D^2 - \dfrac{\left(\sum D\right)^2}{n}}{n-1}}} \tag{15.7}$$

After you have finished your calculations, use Cohen's (1988, 1992) conventions: .2 = small effect, .5 = medium effect, and .8 = large effect.

For our example about comparing driving errors under conditions of no distraction or while text messaging,

$$g = \frac{10.94}{\sqrt{\dfrac{2959.66 - \dfrac{(153.16)^2}{14}}{14 - 1}}} = 1.10$$

Some sources advise using Formula 14.8 to calculate *g* for dependent groups. Do not do this. Formula 14.8 is for independent groups and will give an overestimate of *g.*

What about *r* as a measure of effect size? Recall from Formula 14.9 that

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

r
a correlational measure of effect size

and that for the independent-samples design, *r* is a point-biserial correlation between groups (*X* and *Y*) and scores on a continuous variable. For the dependent-samples design, *r* "is no longer the simple point-biserial correlation, but is instead the correlation between group membership and scores on the dependent variable with indicator variables for the paired individuals partialed out" (Rosnow, Rosenthal, & Rubin, 2000).

15.7 Power

power of a test
the probability of rejecting a false null hypothesis

In Section 13.6, we showed that an increase in sample size increases power by reducing the standard error of the mean. Because the estimated standard error of the mean is $s/\sqrt{n}$, *another way to increase power is to reduce the size of* s. In comparing the means of two groups, *the dependent-samples design makes it possible to reduce the standard error of the difference between the means by controlling the influence of extraneous variables.* We accomplish this effect by pairing observations. If you compare the formula for the standard error of the difference between two correlated means, $s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2 - 2rs_{\overline{X}}s_{\overline{Y}}}$, with that for the difference between two independent means, $s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2}$, you can see the statistical improvement that can be achieved by this method.

15.8 Assumptions When Testing a Hypothesis about the Difference between Two Dependent Means

Unlike the test of the difference between two independent means, the assumption of homogeneity of variance is *not* required for the test between two dependent means. However, hypothesis testing still assumes that our sample has been drawn randomly with replacement and that the sampling distribution of $\overline{X} - \overline{Y}$ (or $\overline{D}$) is normally distributed. Therefore, our choice of sample size is an important design consideration. Remember, in the behavioral sciences most dependent variables are not distributed normally (see Point of Controversy in Chapter 6). When sample size is large (>25), the central limit theorem tells us that the sampling distribution of $\overline{D}$ should approximate the normal curve regardless of what the original population of scores looks like. But what if our sample size is smaller than 25? If samples are small and we have grounds to believe that the parent population is not normal in shape, a "nonparametric" statistical test may be a better alternative than *t* (see Chapter 21).

15.9 Problems with Using the Dependent-Samples Design

In Section 14.12 we discussed the role of random sampling from an experimental point of view. We saw that we can obtain control over the influence of extraneous variables by drawing subjects at random for each treatment condition. We also saw that when random sampling cannot be followed, it is much more difficult to interpret the outcome. Those comments apply to studies using dependent samples as well. In the *matched-subjects* design, we should first draw a random sample from a single population, form matched pairs of subjects, and then randomly assign the treatment condition to the two members of each pair, taking care to do so independently for each pair of subjects. For the *repeated-measures* design, we should decide randomly which treatment the subject will receive first and which will be given second. However, problems can arise when the design of the study departs from these conditions. There are other potential problems as well. Here are three examples.

Study Design Problem 1: Order Effects

order effect when making repeated measurements on the same subjects, exposure to the treatment condition assigned first changes performance on the treatment condition assigned second

When we make repeated measurements on the same subjects, exposure to the treatment condition assigned *first* may change the subject in some way that will affect his or her performance on the treatment condition assigned *second*. An influence of this sort is called an **order effect**. For example, if the two treatment conditions are mild and heavy shocks, a subject receiving the heavy shock first may have a different outlook toward the second trial than one who receives the mild shock first. The outcome then becomes difficult or impossible to interpret.

When an order effect is present *and* it can be assumed that the influence of one treatment on the other is the same as that of the other on the one, the outcome of the experiment may be interpretable, if the order of the treatment condition has been assigned at random. However, the order effect will introduce an additional source of variation in each set of scores, according to the magnitude of its influence. This tends to increase the standard error and, consequently, to decrease the chance of detecting a difference when one exists. The purpose of choosing a dependent-samples design is ordinarily to *reduce* extraneous sources of variation, but the effect may be quite the opposite.

Study Design Problem 2: Repeated-Measures Design with Nonrandom Assignment

If a design uses repeated observations on the same subjects but the assignment of the order of the treatment condition is not random, it may be difficult to interpret the outcome. Any order effect will bias the comparison. In many investigations, it is not possible to avoid this problem. For example, it occurs in studies where the object is to determine what change has taken place in the subjects over a period of time. Is a particular type of psychotherapy effective? The subjects' condition may be evaluated before and after therapy, and the results compared. "Before" and "after" cannot be randomly assigned to a subject. If there is improvement, is it the result of therapy, the person's belief that he or she is *supposed* to feel better, the mere fact that someone paid attention to him or her, or simply a natural recovery that would have taken place without therapy? To be meaningful, the study must include a comparison group, chosen in the same way and treated alike except for receiving the therapy to be evaluated (Bowers & Clum, 1988).

Study Design Problem 3: Sampling Variation in Matched-Subjects Designs

There is an important factor we must consider before using the *matched-subjects* design. Note that the reduction of $\sigma_{\bar{X}-\bar{Y}}$ induced by pairing the observations depends on the value of r, the correlation coefficient induced by pairing. In general, when pairing is on the basis of a variable importantly related to the performance of the subjects, the correlation will be high and the reduction in $\sigma_{\bar{X}-\bar{Y}}$ will consequently be great. A reduction in $\sigma_{\bar{X}-\bar{Y}}$ increases the power of the test. For example, matching subjects on the basis of their intelligence test scores for an experiment on rate of learning will probably be beneficial. However, when we estimate the correlation from the sample, it is rare that the sample coefficient exactly equals the population coefficient. For example, the correlation in the population might actually be zero and yet the sample coefficient might be, say, $+.20$. In practice, this means that we should not match subjects on a variable that "might help," but only on one that we are sure will have a reasonably substantial influence.

15.10 Summary

In this chapter, we introduced procedures for testing for a difference between two dependent groups, where measurements in the two samples are related to each other. There are three ways in which dependent samples may be generated: (1) use of a repeated-measures design, in which the same subjects are used for both conditions of the study; (2) use of a matched-subjects design, in which different subjects are used but are matched by the experimenter on some variable related to performance on the variable being observed; and (3) natural pairings.

The major advantage of the dependent-samples design over the independent-samples design is that an extraneous source of variation (i.e., variation due to subject differences) can be eliminated by pairing observations. For the t test, the effect is reflected in the standard error (the denominator in the formula for t), which is smaller for dependent samples whenever r, the correlation between the set of paired observations, is greater than zero. The greater the correlation, the smaller the standard error, and thus the more likely the experimenter is to notice a true difference between the means of two populations. An alternative formula transforms the t test from a two-sample to a one-sample test by focusing on the distribution of differences, D, between paired X and Y scores. In both cases, the degrees of freedom used to evaluate t for dependent means is $n - 1$, where n is the number of pairs of scores.

Although the dependent-samples design has some advantages, consideration must also be given to some possible problems. In the repeated-measures design, exposure to the treatment condition assigned first may change a subject in some way that will affect his or her performance on the second treatment condition. The repeated-measures design assumes counterbalancing of conditions, but in some types of studies it is not possible to reverse the order of conditions. Interpretation of the outcome in these circumstances can be difficult.

Key Terms, Concepts, and Symbols

dependent-samples design (p. 259)
repeated-measures design (p. 259)
matched-subjects design (p. 259)
matched-pairs investigations (p. 259)
standard error of the difference between dependent means (p. 260)
$\sigma_{\overline{X}-\overline{Y}}$ (p. 260)
estimated standard error (p. 260)
$s_{\overline{X}-\overline{Y}}$ (p. 260)
t for dependent means (pp. 260, 261)
df for t test for dependent samples (p. 260)
difference scores (p. 261)
sampling distribution of differences (p. 261)
D (p. 261)
$\overline{D}$ (p. 261)
estimated standard deviation of the difference scores (pp. 261, 262)
s_D (p. 261)
estimated standard error of the mean of the difference scores (pp. 261, 262)
$s_{\overline{D}}$ (p. 262)
effect size (p. 267)
order effect (269)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In our example problem in Section 15.4, suppose we had used X for the text messaging condition and Y for the no distractions condition. (a) Would it have changed the outcome of the test? (b) Would it have changed the research conclusion? Explain. (c) Why was this a dependent-samples design?

2. Suppose that the scores for the problem in Section 15.4 (Table 15.1) were from two samples drawn independently. Calculate t using Formula 14.6 for two independent means. Compare your answer with that obtained in Table 15.1. Why do the two results differ? If $r = 0$, what effect would it have for the dependent-samples t test?

3. Given the following data from two dependent samples: $\overline{X} = 88.0$, $s_X = 16.0$, $\overline{Y} = 85.0$, $s_Y = 12.0$, $n = 64$, and $r = +.50$. (a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Calculate $s_{\overline{X}-\overline{Y}}$ using Formula 15.2. (c) Complete the test at the .05 and .01 levels of significance and state your conclusions.

4. Given the following pairs of scores from dependent samples:

Pair:	1	2	3	4	5
X:	4	4	6	5	9
Y:	5	2	3	1	6

(a) State formally the hypotheses necessary to conduct a nondirectional test of no difference between the two population means. (b) Calculate t using Formula 15.5. (c) Calculate t using Formula 15.3 (you should get the same answer). (d) Complete the test at the .05 and .01 levels of significance and state your conclusion. (e) Calculate effect size, g.

5. Is reaction time to a red light different from that to a green light? Suppose 20 subjects have been selected at random and given a preliminary test to determine the reaction time of each individual to a white light stimulus. From among this group, 10 pairs of subjects are formed such that the two members of any given pair are equal or nearly equal in speed of reaction to white light. Taking each pair in turn, one member is first tested with a green light as the stimulus and the other with a red light. Which member of each pair receives which stimulus first is determined at random. Reaction times are recorded in milliseconds. The results are given in Data 15A.

Data 15A

PAIR	GREEN LIGHT (X)	RED LIGHT (Y)
1	208	205
2	206	207
3	213	208
4	210	211
5	212	209
6	210	210
7	211	212
8	198	201
9	202	205
10	204	200

(a) State formally the null and alternative hypotheses necessary to conduct a nondirectional test. (b) Complete the test at the .05 level of significance, and state your conclusion. (c) Does it matter which group was designated as X and which as Y? (d) Calculate effect size, g.

6. Dr. Fredericks believes that environment is more important than genetics in influencing intelligence. He locates 12 pairs of identical twins that have been reared apart, one twin in each pair in an enriched environment and the other in an impoverished environment. He administers a standardized IQ test and obtains the results given in Data 15B.

Data 15B

PAIR	ENRICHED	IMPOVERISHED
1	100	102
2	95	93
3	122	116
4	107	110
5	85	75
6	96	100
7	135	124
8	110	110
9	108	100
10	90	88
11	100	90
12	113	108

(a) State formally the null and alternative hypotheses necessary to conduct a nondirectional test. (b) Complete the test at the .05 level of significance, and state your conclusion. (c) Why is this a dependent-samples design? (d) Can Dr. Fredericks conclude that environment is more important than genetics from the results of this test? Explain. (e) Calculate effect size, *g*. (f) Could Dr. Fredericks really have obtained the 12 pairs of twins at random? What effect might this violation of one of the assumptions have on the outcome?

7. A psychologist is interested in the effect of social pressure. She selects a group of 10 male subjects at random from her introductory psychology course. Each subject is asked to estimate by eye the length of a stick under two conditions. In one condition, the subject is placed in the presence of four other persons who, unknown to him, have been instructed to give estimates that are too large. Each of these persons gives his estimate orally to the experimenter before the subject is asked for his estimate. In the other condition, the same procedure is employed, except that the other four persons present are told to give honest estimates. For each subject, the order of the two conditions is determined randomly. Estimates (in inches) made by each subject under both conditions are given in Data 15C.

DATA 15C

SUBJECT	ESTIMATES MADE IN UNBIASED ENVIRONMENT	ESTIMATES MADE IN BIASED ENVIRONMENT
1	38	47
2	35	45
3	37	42
4	40	45
5	45	47
6	34	40
7	40	43
8	36	45
9	43	40
10	41	48

(a) State formally the null and alternative hypotheses necessary to conduct a nondirectional test. (b) Complete the test using $\alpha = .05$, and state your conclusion. (c) Calculate effect size, *g*.

8. A researcher is investigating the effects of a new drug on the ability of rats to learn to navigate a maze. She is worried that differences in activity levels might influence the results, so she gives them a baseline activity test and matches them to form eight pairs. One group is then given the drug and the other group a placebo. Here are the results in terms of number of trials it took each pair to navigate the maze without any errors.

DATA 15D

PAIR	PLACEBO	DRUG
1	6	4
2	4	2
3	5	5
4	7	4
5	3	2
6	5	2
7	4	5
8	6	3

(a) State formally the null and alternative hypotheses necessary to conduct a nondirectional test. (b) Complete the test at the .05 level of significance. (c) Calculate effect size, *g*.

9. The assistant principal at Honey Island elementary school has just learned of a new method of teaching arithmetic to second graders. In order to see whether the new method is better than the traditional method, she selects 30 students at random and assigns them to Mrs. Mathis, who she knows is a good arithmetic teacher. Based on the students' standardized test scores, the assistant principal matches the students in terms of intellectual ability, forming 15 pairs. Mrs. Mathis teaches arithmetic by the traditional method to one member of each pair and uses the new method with the other 15 students. Both groups receive equal amounts of instruction. Which member of each pair receives which instruction method is determined at random. At the end of the year, the students take the California Achievement Test. Their math scores are given in Data I5E.

(a) State formally the null and alternative hypotheses necessary to conduct a nondirectional test. (b) Complete the test using $\alpha = .05$, and state your conclusion. (c) Calculate effect size, *g*.

DATA 15E

PAIR	CAT SCORES UNDER TRADITIONAL METHOD *X*	CAT SCORES UNDER NEW METHOD *Y*
1	78	74
2	55	45
3	95	88
4	57	65
5	60	64
6	80	75
7	50	41
8	83	68
9	90	80
10	70	64
11	50	43
12	80	82
13	48	55
14	65	57
15	85	75

10. For the data in Problem 5, use IBM SPSS to calculate (a) the t statistic and (b) the correlation between X and Y. (c) What is the two-tailed p-value associated with the calculated t? (d) Although IBM SPSS does not report g, use information in the IBM SPSS output to calculate g.

CHAPTER 16

Inference about Correlation Coefficients

When you have finished studying this chapter, you should be able to:

- Understand that Pearson's r varies from sample to sample because of random sampling variation;
- Understand that a correlation coefficient calculated from a sample of scores is statistically significant only if it has a small probability of occurring by chance;
- Test the null hypothesis that there is no relationship between X and Y in the population; and
- Recognize that the same procedures are used for Spearman's r_S when $n \geq 10$.

The Pearson correlation coefficient (see Chapter 7), like other statistics, varies from sample to sample because of chance factors resulting from random sampling variation. In most cases, we know only the sample coefficient, but we would like to know the "true" correlation in the population. In the same manner that we tested hypotheses about population means, we can use hypothesis testing to determine the accuracy of our sample correlation coefficient. Before you begin reading this chapter, make sure you are familiar with the basics of inference and Student's distribution of t (Chapter 12).

16.1 The Random Sampling Distribution of *r*

ρ_{XY}
symbol for Pearson's correlation coefficient calculated from a population of paired scores

r_{XY}
symbol for Pearson's correlation coefficient calculated from a sample of paired scores

random sampling distribution of *r*
a theoretical relative frequency distribution of all the values of *r* that would be obtained by chance from an infinite number of samples of a particular size drawn from a given population

The Pearson correlation coefficient, calculated from a population of paired scores, is $\boldsymbol{\rho_{XY}}$ (Greek letter rho). When the coefficient is calculated from a sample, it is $\boldsymbol{r_{XY}}$. If we drew a sample of given size at random from the population, calculated *r*, returned the sample to the population, and repeated this operation many times, we would see many different values of *r* due to random sampling variation. If we repeated this procedure an *infinite* number of times, the multitude of sample *r*s would form the **random sampling distribution of *r*** for samples of that particular size. The mean of this sampling distribution is approximately ρ, and the standard deviation is σ_r:

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n - 1}}$$

As you can see from the denominator of this formula, σ_r becomes smaller as n increases. Thus, *the values of* r *will vary less from sample to sample when the sample size is large.* Now look carefully at the numerator. The standard deviation becomes smaller as ρ increases. Thus, *the values of* r *will also vary less from sample to sample when the true correlation, ρ, is strong than when it is weak.* Note that because ρ is squared, it is the absolute value that is important, not its sign (positive or negative).

If the sample values of *r* formed a normal distribution, we could solve problems of inference by the methods you learned in earlier chapters. Unfortunately, *the sampling distribution of* r *is not normally distributed.* When $\rho = 0$, the sampling distribution of *r* is symmetrical and nearly normal. But when ρ has a value other than zero, the sampling distribution is skewed because values cannot exceed ±1.00. In fact, the larger the absolute value of ρ, the greater the skewness. Figure 16.1 shows random sampling distributions for $\rho = -.8$, 0, and $+.8$ when $n = 8$. This graph helps explain why the distribution is skewed when the absolute value of *r* is high. When $\rho = +.8$, for example, it would not be possible for us to obtain a sample coefficient more than .2 point higher ($r = +1.0$), but random sampling could produce a coefficient substantially lower than .2 point below $r = +.8$ (to a lower limit of -1.0).

The degree of skewness of the sampling distribution of *r* is also a function of sample size. When $\rho \neq 0$, the smaller the size of the sample, the greater the skewness of the sampling distribution.

Because the distribution of sample *r*s is not normal, we must seek alternative solutions for making inferences about population correlation coefficients. For some problems, the *t*

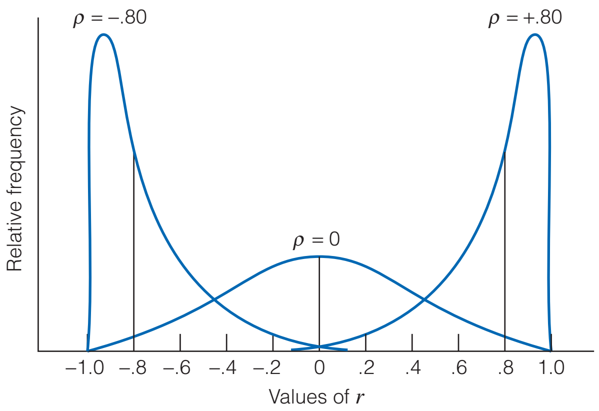


FIGURE 16.1 Random sampling distributions of *r* for three values of ρ; $n = 8$.

distribution will give us an appropriate model. For others, transformation of r to a variable that is (approximately) normally distributed offers the best solution.

16.2 Testing the Hypothesis that $\rho = 0$

$\rho = 0$
the usual null hypothesis in inference about correlation

Without doubt, the most frequent question of inference about correlation is whether there is *any* relationship at all between two variables. Suppose, for example, that we are developing a test with the hope of predicting success in the job of insurance salesperson. We want to know, of course, whether the test has *any* relationship to actual job success. To determine this, we draw a random sample of 25 salespersons and find $r = +.30$ between test performance and job performance rating. It may be that the true correlation (ρ) is zero and that we obtained this value of r simply as a matter of chance (random sampling variation). To check on this possibility, we will test the null hypothesis that $\rho = 0$ against the alternative that it does not. (Before doing this, however, we should always examine the scatter diagram to see whether the cluster of dots is in agreement with our calculated value of r—see Chapter 7.)

Sir Ronald Fisher showed that when $\rho = 0$, the value obtained by the following formula is distributed as Student's t test with $n - 2$ degrees of freedom:

$t = \dfrac{r}{\sqrt{(1 - r^2)/(n - 2)}}$
formula for t for testing the null hypothesis $\rho = 0$

t REQUIRED FOR TESTING THE
NULL HYPOTHESIS THAT $\rho = 0$

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} \qquad (16.1)$$

where: r = the sample coefficient
n = the number of *pairs* of scores in the sample.

To test the null hypothesis that $\rho = 0$, we could use Formula 16.1 and refer to the values of t found in Table D in Appendix D.

This method is quite general. We can use it for a variety of levels of significance and for one- or two-tailed tests. (We use a two-tailed test for either a positive or a negative correlation; a one-tailed test for one or the other, but not both.) However, if we wish to conduct tests at the usual 5% or 1% significance level, Table G in Appendix D eliminates the chore of computing t. This table gives the critical value of r required to reach significance. (The rs are expressed as absolute values; for tests of negative correlations, insert a minus sign.) To use this table, find the appropriate value of *df* and read the critical value of r for the desired significance level. If r_{calc} exceeds the tabulated value, we reject the null hypothesis of no correlation; otherwise, we do not. For our example, if we conduct a two-tailed test at $\alpha = .05$, and if $r_{\text{calc}} = .30$ based on $n = 25$, we first determine the degrees of freedom:

$$df = n - 2 = 25 - 2 = 23$$

We then enter Table G and look down the left column until we locate this number of degrees of freedom. We then look across to find the entry in the column for a two-tailed test at $\alpha = .05$. This is the critical value of r:

$$r_{\text{crit}} = .396$$

Because our obtained value of $r = +.30$ is less than this critical value, we cannot reject the null hypothesis of no correlation. Our r_{calc} is one of many that could have occurred by chance when the null hypothesis of no correlation is true. Figure 16.2 illustrates this situation.

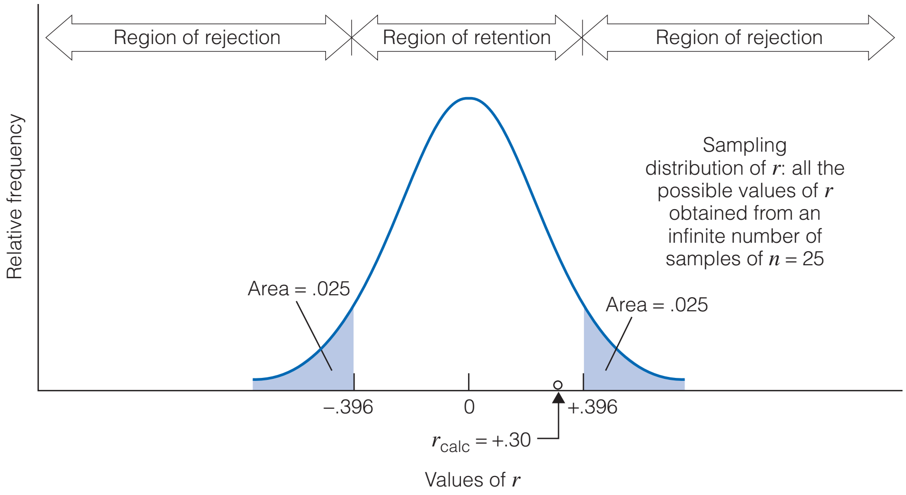


FIGURE 16.2 Testing the hypothesis that $\rho = 0$ when $r = +.30$ and $df = 23$.

What if our r_{calc} proves to be significant? Be careful, a significant r does not mean that the association is important. The expression "significant correlation" means only that H_0: $\rho = 0$ has been tested and rejected, and "nonsignificant correlation" means only that H_0: $\rho = 0$ has been tested and retained—nothing more, nothing less. With large samples, a very small (and perhaps unimportant) r may prove to be "statistically significant." Remember from Section 13.1 that the question of whether a result is important or not depends on the research question, not the statistical question.

If you are calculating r with IBM SPSS, the final printout not only gives you r, but also the level of significance for a two-tailed test. For example, look again at Table 7.6.

16.3 Fisher's z' Transformation

Fisher's z' a logarithmic function of r, used because the sampling distribution of r is not normally distributed when $\rho \neq 0$

Although we can use the technique described in the previous section to test the hypothesis that $\rho = 0$, we cannot use it to answer other questions about correlation. The fundamental problem is that the sampling distribution of r is not normally distributed when ρ is not equal to zero. Moreover, the degree of skewness varies with the magnitude of ρ. Fisher derived a logarithmic function of r, which we will call **Fisher's z'** (z prime), that has certain desirable properties:

1. The sampling distribution of z' is approximately normal, irrespective of the value of ρ.
2. The standard error of z', unlike the standard error of r, is essentially independent of the value of ρ.

Thus, we can translate r to z' and then use the familiar properties of the normal curve to conduct the evaluation. The outcome is the same as though inference were conducted directly in terms of r.

The formula for z' is

FORMULA FOR TRANSLATING
r TO FISHER'S z'

$$z' = \tfrac{1}{2}[\log_e(1 + r) - \log_e(1 - r)] \quad (16.2)$$

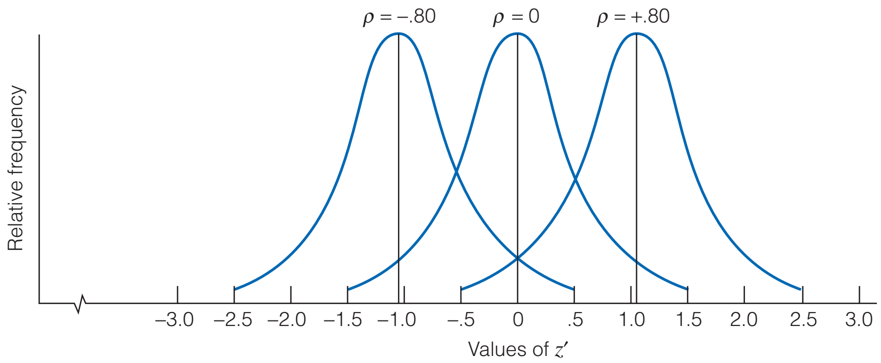


FIGURE 16.3 Random sampling distributions of z' corresponding to three values of ρ; $n = 8$.

Note that the value of z' depends *wholly* on the value of r. The z' transformation is simply a nonlinear mathematical reformulation that changes the scale of r.[1] Note, also, that *Fisher's* z′ *is not a* z *score, nor is it related to the* z *score in any way.* The similarity in symbolism is unfortunate; another symbol would have been better. Fisher originally called his statistic z, which is completely unacceptable because of the wide use of that symbol for a standard score. Usage has perpetuated Fisher's symbolism as closely as possible without creating utter confusion.

You can best appreciate the improvement in scale that the use of z' provides through an illustration. In Figure 16.3, we show the sampling distributions of z' that correspond to $\rho = -.8$, $\rho = 0$, and $\rho = +.8$ when $n = 8$. Compare these distributions with those of r for the same values of ρ (Figure 16.1). Note that the sampling distributions of z' are essentially similar in shape and variability under these three conditions, whereas those of r are not.

It would be awkward to use Formula 16.2 to translate an r to z', or vice versa. Fortunately, Table H in Appendix D makes the conversion easy. In this table, if r is positive, read the value of z' as positive, and if r is negative, read z' as negative.

Give this table a few moments of study; it reveals what is accomplished by Fisher's transformation. As r increases from 0 to 1, a discrepancy between r and z' begins to appear, with z' being larger than its parent, r. Note, however, that when r is 0, z' is also 0, and that the two values are essentially equivalent up to, say, $r = .25$. The discrepancy between r and z' increases more rapidly as the value of r gets higher. What is accomplished is that the scale of r (which ranges from -1.00 through 0 to $+1.00$) is stretched, very little at the center but increasingly more so at the extremes. Note that the original scale of r becomes, in terms of z', a scale ranging from approximately -3 to $+3$ (or even farther, if r is allowed to exceed $\pm.995$).

For most uses of the z' transformation, keep in mind that we can obtain reasonable results unless the sample size (n) is very small (because of considerable random sampling variation) or ρ is very high (for the reasons discussed in the preceding paragraph). Actually, we have no justification for finding Pearson's r among very small samples, nor do we ordinarily encounter correlation coefficients much above .9. Thus, the limitations are more academic than practical.

---

[1] More generally in mathematics, this transformation is known as the inverse hyperbolic tangent.

16.4 Strength of Relationship

coefficient of determination (r^2) the proportion of the *Y* variance that is associated with differences in *X*

Suppose your obtained value of r proves to be statistically significant; what about the strength of the relationship? We have already introduced ways of interpreting r in Chapter 9. Be sure to look at Section 9.2 (the slope of the regression line) and Section 9.5 (the coefficient of determination). Recall that r^2, the coefficient of determination, gives the proportion of the Y variance that is associated with differences in X.

16.5 A Note about Assumptions

From earlier chapters, you may recall that no assumption about the shape of the bivariate distribution is required when the correlation coefficient is used purely as a descriptive index. However, *all the procedures for inference about coefficients described in this chapter are based on the assumptions that there is independence among the pairs of scores and that the population of the pairs of scores has a normal bivariate distribution.* The latter assumption implies that X is normally distributed, that Y is normally distributed, and that the relation between X and Y is linear. If we are not dealing with a normal bivariate population, then the procedures for inference we have discussed here will yield only approximate results. If there is reason to believe that the population of the pairs of scores varies considerably from a normal bivariate distribution, you might consider nonlinear transformations (to a normal distribution) before conducting formal hypothesis testing (Fox, 2008).

16.6 Inference When Using Spearman's r_S

r_S symbol for Spearman's rank-order correlation coefficient

In Chapter 7 (Section 7.9) we introduced an alternative to Pearson's r when the paired scores are both in the form of ranks and n is rather small. If we choose to use Spearman's r_S, how do we test the hypothesis that the population value is zero?

Look again at the example problem in Chapter 7 (Section 7.9). The calculated value of r_S was $+.38$. Can the instructor conclude that those who turn their papers in earlier tend to earn poorer scores? Exact procedures have been developed for testing the hypothesis of no correlation for very small samples. We can obtain good results for $n \geq 10$ by finding the critical values required for significance for $df = n - 2$ in Table G in Appendix D. This is the same table we use to determine the significance of Pearson's r. For the present problem, Table G shows that for $df = n - 2 = 13$, the instructor would need a coefficient of $+.514$ or larger to reject the hypothesis of no correlation for a two-tailed test at $\alpha = .05$. The obtained coefficient of .38 fails to meet this criterion, so the instructor retains the null hypothesis of no correlation in the population.

16.7 Summary

Like other statistics, the sample correlation coefficient, r, is subject to sampling error. In this chapter, we introduced statistical inference techniques to test the hypothesis that there is no linear association between two variables (H_0: $\rho = 0$). The test is a t test with $n - 2$ degrees of freedom, where n is the number of pairs of observations. For Spearman's r_S, we use the same procedure when $n \geq 10$.

Whether or not significance is reached depends on the size of the sample. There is generally a considerable amount of sampling variation with small samples, making it difficult to obtain statistical significance. For large samples, the value of r will be fairly accurate, and statistical significance may be achieved even with a very low degree of association between two variables. We must be careful, therefore, to distinguish between statistical significance and practical significance when dealing with correlations.

Key Terms, Concepts, and Symbols

ρ_{XY} (p. 275)

r_{XY} (p. 275)

random sampling distribution of r (p. 275)

$\rho = 0$ (p. 276)

significant correlation (p. 277)

Fisher's z' (p. 277)

r^2 (p. 279)

coefficient of determination (p. 279)

Spearman's r_S (p. 279)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. What sample value of r would we need for statistical significance if (a) $n = 5$ and $\alpha = .05$ for a two-tailed test? (b) $n = 10$ and $\alpha = .05$ for a two-tailed test? (c) $n = 25$ and $\alpha = .05$ for a two-tailed test? (d) $n = 25$ and $\alpha = .01$ for a two-tailed test? (e) $n = 25$ and $\alpha = .05$ for a one-tailed test? (f) $n = 100$ and $\alpha = .05$ for a two-tailed test? (g) $n = 1000$ and $\alpha = .05$ for a two-tailed test?

2. Given: H_0: $\rho = 0$; H_A: $\rho \neq 0$; $r = +.35$; $n = 24$. Test the null hypothesis at $\alpha = .10$, .05, and .01.

3. Using Table G in Appendix D, test the null hypothesis that $\rho = 0$ when (a) $r = +.40$, $n = 28$, H_A: $\rho \neq 0$, and $\alpha = .05$. (b) $r = -.55$, $n = 18$, H_A: $\rho \neq 0$, and $\alpha = .01$. (c) $r = -.40$, $n = 22$, H_A: $\rho < 0$, and $\alpha = .05$. (d) $r = +.45$, $n = 25$, H_A: $\rho < 0$, and $\alpha = .05$. Show the critical value(s) of r for each problem.

4. The director of human resources at an automobile assembly plant is not satisfied with the performance of recently hired employees. To improve matters, he decides to use an employment test to screen applicants. A group of 14 new applicants is given the test and hired. Three months later they receive their first on-the-job ratings by supervisors. The correlation between test scores and job ratings is found to be $+.41$. Can the director of human resources claim that the test successfully predicts job performance? Use $\alpha = .05$ and a two-tailed test.

5. In a particular study, it is reported that $r = +.75$ between predictor and criterion. You find that this value was obtained on a sample of six cases. Comment.

6. In another study, it is reported that $r = +.12$ between predictor and criterion. This correlation is based on 500 cases. Is it significantly different from zero? Comment.

7. (a) If all we know about a study is that the correlation between two variables was "not significant" for a sample of 1,000 cases, can we adequately evaluate the results without knowing r? Explain. (b) If the result had been "significant," can we adequately evaluate the results without knowing r? Explain.

8. Look again at Problem 8 in Chapter 7. The calculated value of r_S was $+.60$. (a) Is r_S significantly different from zero according to a two-tailed test at the 5% significance level? Show the basis for your answer. (b) Do you have any qualifications for your answer to Problem 5? Explain.

9. Look again at Problem 10 in Chapter 7. The calculated value of r_S was $+.83$. Is r_S significantly different from zero according to a two-tailed test at the 5% significance level?

CHAPTER 17

An Alternative to Hypothesis Testing: Confidence Intervals

When you have finished studying this chapter, you should be able to:

- Understand the merits of confidence intervals compared to dichotomous hypothesis testing;
- Understand what is meant by the terms "95% confidence interval" and "99% confidence interval";
- Calculate a confidence interval for μ_X or $\mu_X - \mu_Y$ with use of original scores and in terms of standard deviations of the variable;
- Calculate sample size required for estimates of μ_X or $\mu_X - \mu_Y$; and
- Calculate a confidence interval for the population correlation coefficient, ρ.

> Interval estimates should be given [when presenting results] for any effect sizes involving principal outcomes. . . . Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies. (Wilkinson and the Task Force on Statistical Inference. APA Board of Scientific Affairs, 1999)

The inferential technique of testing the null hypothesis is credited to Ronald A. Fisher, "the father of modern statistics." It is a mechanical methodology resulting in one of only two possible decisions: Either we reject the null hypothesis or we retain it.

In the real world, the null hypothesis is almost always false; it is almost certain that the population parameter under test is not exactly equal to the value stated in the null hypothesis. The question researchers should ask themselves, according to some critics (Cohen, 1990; Rosnow & Rosenthal, 1989b), is by how much the null hypothesis is incorrect, not whether a result is "statistically significant" or not. Interestingly, Fisher opposed the idea of an alternative hypothesis. This was the creation of Jerzy Neyman and Egon Pearson (1928a, b), whose views Fisher vehemently opposed (Gigerenzer, 1987). Nevertheless, it became standard practice that when rejecting the null hypothesis, one accepts the alternative. But with **dichotomous decision making**, if one fails to reject the null hypothesis, one cannot conclude anything about it at all. Only a "significant result" allows a definite conclusion.

dichotomous decision making null hypothesis significance tests can result in one of only two possible decisions: either we reject H_0 or we retain it

This methodology developed from agricultural experiments, where research required making decisions (for example, whether a particular fertilizer was effective) (Gigerenzer, 1987). Decision-oriented research may be appropriate in some fields, but the research questions asked in the behavioral sciences are often quite different (Rosnow & Rosenthal, 1989b). Nevertheless, most research is conducted within the Fisherian dichotomous decision-making scheme.

Compounding the problem is the sanctity that the behavioral sciences have bestowed on the .05 level of significance. Is there a basis for the tradition of using .05 as the sharp line of division between a significant and a nonsignificant result? In other fields, α levels other than .05 are traditional. The acclaimed physicist Enrico Fermi, for example, believed the operational definition of a "miracle" to be $p = .10$ (Polanyi, 1961). The problem is not whether the behavioral sciences should use .05 or some other value as the traditional α level, but dichotomous decision making itself. There is, in fact, no ontological basis for it:

***p*-value** the probability, when H_0 is true, of observing a sample mean as deviant as or more deviant (in the direction specified in H_A) than the obtained value of $\overline{X}$

> Surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of *p*? (Rosnow & Rosenthal, 1989b)

The controversy about dichotomous hypothesis testing has never been greater than it is today. In fact, some statisticians and psychologists have called for a ban of significance tests altogether (e.g., Cohen, 1994; Cumming, 2008; Hunter, 1997; Kehle et al., 2007; Loftus, 1996; Schmidt, 1996; see comment section in *American Psychologist*, 53(7), 1998). Others continue to defend its use (e.g., Abelson, 1997; Chow, 1998; Hagen, 1997; Wainer, 1999). Still others argue for a balanced approach, finding merit (and fault) with both arguments (e.g., Krueger, 2001).

The American Psychological Association's Task Force on Statistical Inference did not ban null hypothesis significance testing, but strongly recommended that in addition to *p*-values, researchers provide effect size and confidence intervals (see the opening quote). In this chapter, you will learn about using confidence intervals as an alternative to dichotomous hypothesis testing.

17.1 Examples of Estimation

A Harris poll of nearly 2,000 registered voters conducted over the last 3 days of the 1992 presidential election found that 44% preferred Bill Clinton, 38% preferred George H. W. Bush, and 17% preferred Ross Perot, with only 1% still undecided. In five other major media polls, Mr. Clinton's support was identical in four of them—44%—and in a fifth it was 43%. Mr. Bush's support ranged from 35% to 38% and Mr. Perot's from 14% to 17% (*New York Times,* November 3, p. A9). In all surveys, the margin of error was reported as 3%. According to

The New York Times, which participated in a poll with *CBS News*, this meant that "in 19 cases out of 20 the results based on such samples will differ by no more than three percentage points in either direction from what would have been obtained by seeking out all American adults." On November 3, over 100 million people cast their votes. The final results were 43% for Clinton, 38% for Bush, and 19% for Perot (most of the undecided voters apparently voted for Perot). The political surveys, which questioned only a small fraction of the registered voters, had been accurate in their projections.

The 2008 presidential election posed some novel problems for pollsters. Among them was the question of whether or not likely voters would be honest in their answers about voting for an African-American candidate. Yet, in the end, most polls were accurate. On election day, Barack Obama beat John McCain 52% to 46% and most polls were within 2% of these numbers (Bialik, 2008).

The Harris and major media surveys are examples of estimation. In estimation, unlike hypothesis testing, no value is specified in advance. The question is, "What *is* the value of the population parameter?" For some kinds of research questions, such as those ascertaining what percentage of voters prefer a particular candidate, hypothesis testing is useless; no specific hypothesis presents itself. On the other hand, estimation procedures are exactly suited to such problems. Although there are some research questions that cannot be answered by hypothesis testing but can be approached through estimation, the reverse is not true. *Any problem for which hypothesis testing might be useful can alternatively be approached through estimation.* In this chapter, we will concentrate on the estimation of means. (The previous example was about percentages; no means were involved.)

point estimates estimations of a single value of a population parameter

An estimation may take two forms. Sometimes we desire a single value as an estimate of the population value. We call such estimates **point estimates**. What percentage of voters will vote for candidate X? What is the mean aptitude test score of applicants for admission to Spartan University? If it is impractical to find the percentage in the entire population (as in the first example) or the mean of the entire population (as in the second example), we can make an estimate of the population characteristic from a random sample.

interval estimates estimations of a range of values of a population parameter

Point estimates are rarely stated by themselves. Because of the chance factors involved in random sampling, any point estimate is likely to be in error. But by how much will it differ from the population parameter? For **interval estimates**, we calculate a range of values (an interval) within which it appears reasonable that the population parameter lies. In the question about candidate X, a point estimate might state that 49% of the population of voters favor him or her. If we made an interval estimate, the outcome might state that we are 95% confident that no less than 46% and no more than 52% of the voters (in the population) favor him or her.

17.2 Confidence Intervals for μ_X

Remember from Chapter 6 that in a normal distribution, 95% of the scores are no farther away from the mean than 1.96 standard deviations. Similarly, if the sampling distribution of $\overline{X}$ is normally distributed, 95% of the sample means are no farther away from μ_X than 1.96 standard errors of the mean ($\sigma_{\overline{X}}$). Now, *if 95% of the sample means are no farther away from μ_X than $1.96\sigma_{\overline{X}}$, it is equally true that for 95% of the sample means, μ_X is no farther away than $1.96\sigma_{\overline{X}}$.* This fact makes it possible for us to construct an interval estimate of μ_X. Suppose for each sample mean the statement is made that μ_X lies somewhere within the range $\overline{X} \pm 1.96\sigma_{\overline{X}}$. For 95% of the sample means, this statement would be correct, and for 5% it would not. In drawing samples at random, the probability is therefore .95 that an interval estimate constructed according to the rule

$$\overline{X} \pm 1.96\sigma_{\overline{X}}$$

will include μ_X within the interval.

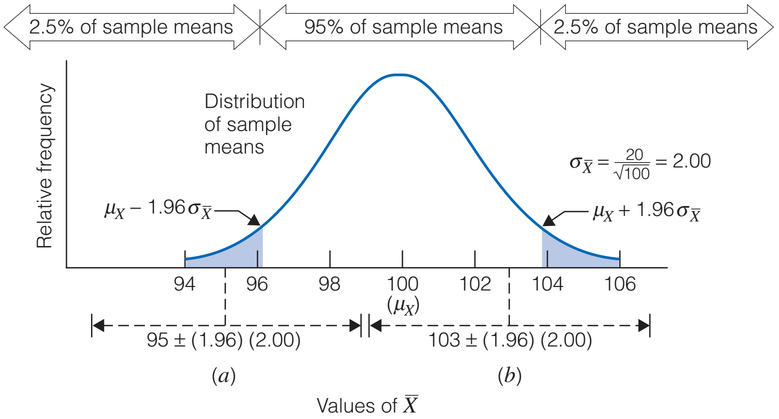


FIGURE 17.1 Distribution of sample means based on $n = 100$, drawn from a population characterized by $\mu_X = 100$ and $\sigma_X = 20$. (*a*) Confidence interval of μ_X constructed about $\bar{X} = 95$; estimate does not cover μ_X. (*b*) Confidence interval of μ_X constructed about $\bar{X} = 103$; estimate covers μ_X.

We illustrate the application of this rule in Figure 17.1. Suppose we draw random samples of size 100 from a normally distributed population of IQ scores in which $\mu_X = 100$ and $\sigma_X = 20$. In this example, $\sigma_{\bar{X}}$ is $20/\sqrt{100} = 2.00$. If we draw a sample whose mean is 103, we would claim that μ_X lies somewhere within the interval $103 \pm (1.96)(2.00)$ or between 99.08 and 106.92. This interval is shown in (*b*) in Figure 17.1, and we see that the claim is correct. On the other hand, if $\bar{X}$ had been 95, we would claim that μ_X lies within the interval $95 \pm (1.96)(2.00)$ or between 91.08 and 98.92. This interval is shown in (*a*); the claim is wrong because 95 is one of the 5% of sample means that lie farther from μ_X than $1.96\sigma_{\bar{X}}$.

A 95% *confidence interval* for μ_X is an interval so constructed that 95% of such intervals include μ_X. We may wish to be even more sure that our estimate includes μ_X. Suppose that we prefer a probability of .99 instead of .95. Because 99% of sample means fall between $\mu_X \pm 2.58\sigma_{\bar{X}}$, we may make an estimate for which the probability is .99 that the interval will cover μ_X by the rule

$$\bar{X} \pm 2.58\sigma_{\bar{X}}$$

Notice that the result is a wider interval than when a probability of .95 was used. Although the two levels of probability, .95 and .99, are the ones commonly used in interval estimation, it is possible to construct an interval estimate according to any desired level of probability.

The procedures that we have described require knowledge of $\sigma_{\bar{X}}$, which in turn requires that we know σ_X. This normally will not be the case. Thus, as usual, we must substitute $s_{\bar{X}} = s_X/\sqrt{n}$ as an estimate of $\sigma_{\bar{X}}$ (see Section 12.9). The manner in which we construct a confidence interval is identical except that we substitute the appropriate value of t for z:

$s_{\bar{X}}$
symbol for the estimated standard error of the mean

t_p
symbol for the magnitude of t for which the probability is p of obtaining a value of $\bar{X}$ so deviant or more so

RULE FOR CONSTRUCTING A CONFIDENCE INTERVAL FOR μ_X WHEN σ_X IS UNKNOWN

$$\bar{X} \pm t_p s_{\bar{X}} \tag{17.1}$$

where: $\bar{X}$ = the sample mean, obtained by random sampling
$s_{\bar{X}}$ = the estimate of the standard error of the mean
t_p = the magnitude of t for which the probability is p of obtaining a value so deviant or more so (in either direction); $df = n - 1$, the number of degrees of freedom associated with s_X.

Problem

Suppose that the following results on a mathematics proficiency exam are obtained for a random sample of 25 college freshmen at Spartan University: $\overline{X} = 85.0$ and $s_X = 15.0$. Construct the 95% confidence interval for μ_X.

Solution

$\overline{X} \pm t_p s_{\overline{X}}$
formula for constructing a confidence interval for μ_X

We first calculate the estimated standard error of the mean:

$$s_{\overline{X}} = \frac{s_X}{\sqrt{n}} = \frac{15.0}{\sqrt{25}} = 3.00$$

The number of degrees of freedom associated with s_X is $n - 1$, or 24. From Table D in Appendix D, the appropriate value of t is 2.064. The limits are therefore

$$\overline{X} \pm t_{.05} s_{\overline{X}}$$

Lower Limit	Upper Limit
$= 85.0 - (2.064)(3.00)$	$= 85.0 + (2.064)(3.00)$
$= 78.8$	$= 91.2$

confidence interval
the limits of the interval estimate

confidence coefficient
the degree of confidence

C
symbol for the confidence coefficient

When we construct intervals according to the rule used in the preceding problem, it is proper to say that the probability is .95 that an interval so constructed will include μ_X. However, *once we establish the specific limits for a given set of data, the interval thus obtained either does or does not cover* μ_X. It is *not* proper to say that the probability is .95 that μ_X lies within the interval. The probability at this stage is either 1.00 or 0 that the interval covers μ_X; we do not know which. Consequently, it is usual to substitute the term *confidence* for *probability* in speaking of a *specific* interval. The limits of the interval are referred to as a **confidence interval** and the statement of degree of confidence is called the **confidence coefficient**, ***C***. In our example,

$$C(78.8 \leq \mu_X \leq 91.2) = .95$$

We translate the statement as follows: We are 95% confident that μ_X falls between 78.8 and 91.2.

But what does it mean to say that we are "95% confident"? We do not know whether the *particular* interval covers μ_X, but when intervals are constructed according to the rule, 95 of every 100 of them (on the average) will include μ_X. Remember that *it is the interval that varies from estimate to estimate and not the value of* μ_X. Because μ_X is a fixed value, it does not vary. Figure 17.2

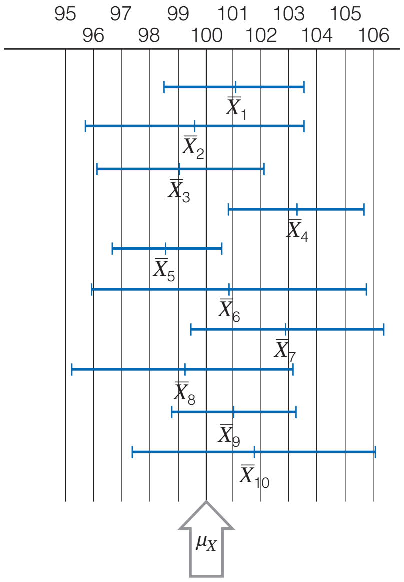


FIGURE 17.2 Interval estimates of μ_X constructed from means of several random samples; σ_X unknown.

shows some of the intervals that might result if we selected samples at random and constructed intervals around the sample means by the procedures just described. We may expect that 95% of such estimates will include μ_X within their range when t_P is selected to agree with $C = .95$.

In Figure 17.2, notice that as we move from sample 1 to sample 10, the *locations* of the intervals vary because the sample means vary. Notice also that the *widths* of the intervals differ, too. This is because interval width depends in part on the varying estimates of the population standard deviation obtained from the several samples.

Recall that s_X (and consequently, $s_{\overline{X}}$) is affected by sample size. Thus, *for a given confidence coefficient, a small sample results in a wide confidence interval and a large sample results in a narrower one.* We explain how to choose sample size so that the resulting interval is of a desired width in Section 17.8.

17.3 The Relation between Confidence Intervals and Hypothesis Testing

Confidence intervals and hypothesis testing are two sides of the same coin. For most population parameters, *the confidence interval contains all values of* H_0 *that would be retained had they been tested using* $\alpha = 1 - C$ (for a *nondirectional* alternative hypothesis). If the value specified in H_0 falls within the interval, we would retain H_0 in hypothesis testing, whereas if that value falls outside the interval, we would reject it.

Consider, for example, a problem in which $\overline{X} = 112$, $s_{\overline{X}} = 5$, and $n = 25$. When the null hypothesis H_0: $\mu_X = 100$ is tested, $t = 12/5 = +2.4$; the difference is significant at the .05 level ($t_{\text{crit}} = \pm 2.064$ for 24 *df*). If we calculate a confidence interval from exactly the same data, we get $C(+101.68 \leq \mu_X \leq +122.32) = .95$. This interval estimate says that the value of μ_X exceeds 100 by 1.68 to 22.32 points. Notice that the interval does *not* include the possibility that $\mu_X = 100$, the same message conveyed by the null hypothesis. In addition, the confidence interval informs us what the true difference *could* be.

Consider a second example: If a confidence interval reads $C(-16.70 \leq \mu_X \leq +6.70) = .99$, zero is among the possible values that the difference might take. If instead of constructing a confidence interval, we had tested the null hypothesis of no difference at $\alpha = .01$, it would have produced a t smaller than the value needed to reject the hypothesis.

Two notes of caution need to be addressed before we proceed. First, some statisticians have suggested that we use confidence intervals to test hypotheses (in the manner introduced in the two previous paragraphs). Keep in mind, however, that when used in that way, confidence intervals have the same Type I and Type II error rates as does null hypothesis significance testing (Harris, 1997). Second, a survey of authors of journal articles found that many do not really understand the difference between confidence intervals and standard error bars (Belia et al., 2005). Be careful; they are not the same!

17.4 The Advantages of Confidence Intervals

The behavioral sciences employ hypothesis testing much more often than confidence intervals, but in many studies confidence intervals are superior (Cohen, 1994; Cumming et al., 2007; Hunter, 1997; Schmidt, 1996). Let us compare the two approaches with particular attention to inference about means.

1. The final quantitative output of a confidence interval is a statement about the parameter of interest. In hypothesis testing, the statement is about a derived score, such as z or t, or about a probability, *p*. In both forms of inference, the question is about the parameter. In estimation, unlike hypothesis testing, it is right before our eyes.

2. A confidence interval straightforwardly shows the influence of random sampling variation and, in particular, sample size. Remember that for a given level of confidence, a small sample results in a wide confidence interval and a large sample results in a narrower one. The width of the interval gives the investigator a direct indication of whether the estimate is precise enough for the purpose at hand. For hypothesis testing, significance is subject to *two* influences that cannot be untangled without further study: (1) the difference between what was hypothesized and what is true and (2) the amount of sampling variation present. For example, a large value of t could occur in a study where the true difference was small but the sample was large, or where the true difference was large and the sample was small. Unless we look at the descriptors and n, we do not know which one it is.
3. In hypothesis testing, it is easy to confuse a statistically significant difference with an important one (Section 13.1). Essentially, this problem disappears when we use a confidence interval. Suppose that for a sample of IQ scores, $\overline{X} = 103$, $s_X = 20$, and $n = 1600$. If we test the hypothesis that $\mu_X = 100$, we obtain a t of approximately $+6$ with a corresponding probability $<.000000001$. Impressive! However, the 95% confidence interval of μ_X is $C(102 \leq \mu_X \leq 104) = .95$, which brings us back to reality. It is up to us to decide whether this result is important.
4. The outcome of testing a hypothesis is to declare that the *one* specific condition stated in the null hypothesis could be true (H_0 is retained) or is unlikely to be true (H_0 is rejected). A confidence interval, on the other hand, emphasizes the existence of a range of values that might characterize the parameter in question. Where we can, it seems desirable to recognize the variety of possible outcomes rather than develop a "yes–no" mindset toward one particular possibility.

17.5 Random Sampling and Generalizing Results

A word of caution is necessary before we proceed. There have been some noted mistakes when people have used estimation techniques. In the 1948 presidential election, the *Chicago Daily Tribune* published a paper with the headline "Dewey Defeats Truman" (Figure 17.3), when in fact Truman won. They based the story (rather, a prediction) on opinion polls taken before the election. In the 2000 presidential election, the last *New York Times/CBS News* poll (taken 3–5 days before the election) indicated that George W. Bush would beat Al Gore by five percentage points. On election day, 48.1% voted for Bush and 48.2% for Gore.

Do such results condemn the use of estimation techniques? The answer is "no." In Sections 13.12 and 14.12 you learned that *we may generalize results only to the population from which the observations may be considered to be a random sample.* Just as with hypothesis testing, the techniques you learn about in this chapter assume random sampling. Otherwise, you run into the same problem as the *Literary Digest* did when they attempted to predict the outcome of the 1936 presidential election (Section 11.1). Even with random sampling, *we cannot properly generalize beyond the exact set of conditions at the time* (Sections 13.12 and 14.12). The last polls taken before a presidential election are usually taken a few days *before* the election. With random sampling, the results may be quite accurate for that time period, but over the last few days the undecided voters make a decision, and some of the others might change their decision. Thus, caution about generalizing results holds true for confidence intervals established around sample means as well.

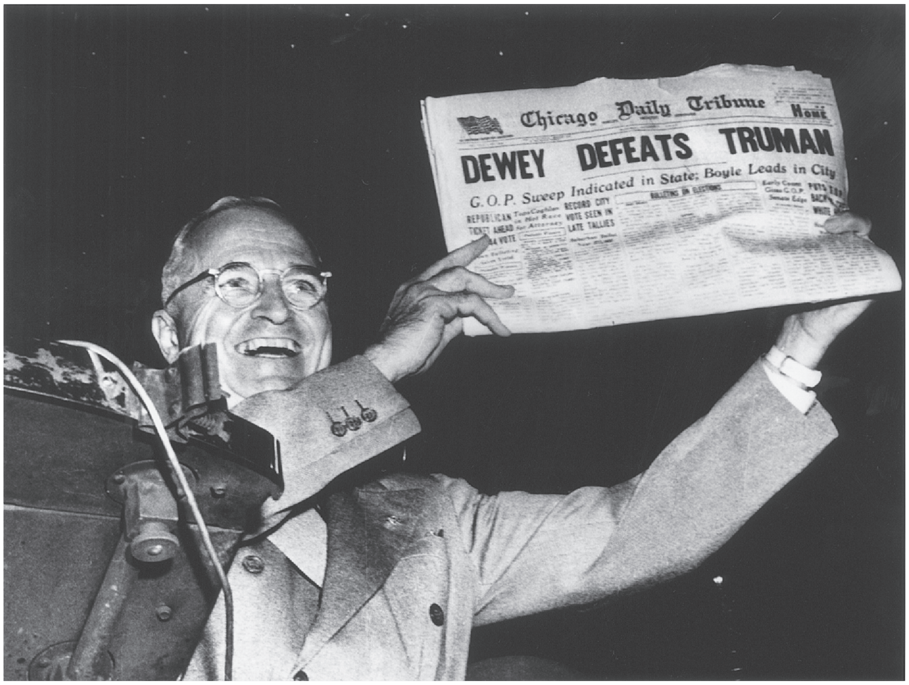

Chicago Daily Tribune
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in City

FIGURE 17.3 In 1948, the *Chicago Daily Tribune* declared that Dewey had defeated Truman. This was actually a prediction based on surveys taken before the election. This is a good example of why we cannot properly generalize sample results beyond the exact set of conditions at the time. © Bettmann/CORBIS

17.6 Evaluating a Confidence Interval

In the example we used to demonstrate Formula 17.1, $\overline{X} = 85.0$, $s_X = 15.0$, and $n = 25$. The 95% confidence interval for μ_X was: $C(78.8 \leq \mu_X \leq 91.2) = .95$. According to this estimate, μ_X is somewhere between 78.8 and 91.2. A statement of this kind has meaning when we are acquainted with the variable. For example, college administrators understand reasonably well what a College Entrance Examination Board score of 400 or 750 means. But many times we are not that familiar with the variable under study. For example, are you familiar with measuring change in the electrical resistance of the skin under various kinds of emotional stimulation? In cases like this, a confidence interval expressed in terms of the variable measured may not be very meaningful.

One way to add meaning to a confidence interval is to interpret the interval limits in terms of number of standard deviations of the variable rather than in terms of raw-score points. The interval estimate cited at the opening of this section was constructed by calculating $85.0 \pm (2.064)(3.0)$ or 85.0 ± 6.19. In effect, then, the confidence interval as constructed states that μ_X is no farther away from $\overline{X}$ than 6.19 raw-score points. Let us reexpress this difference in terms of the number of standard deviations of the variable. The estimated standard deviation is 15.0, so the difference becomes $6.19/15.0 = .41$. Reinterpreted, the interval estimate now states that μ_X is no farther away from $\overline{X}$ than .41 of a standard deviation, which for many purposes would be considered fairly close.

One advantage of expressing a confidence interval in this way is that it compensates for the fact that the importance of a given interscore distance depends on the size of the standard deviation of the variable. For example, a difference of .3 point is minuscule for GRE (Graduate Record Examination) scores, which have a mean of 500 and a standard deviation of 100, but it is quite substantial for grade point averages. By comparing the difference obtained to the standard deviation of the variable, confidence intervals in both these dimensions are placed on common ground.

d
the difference between $\overline{X}$ and the outer limits of the interval estimate, expressed in terms of the number of standard deviations of the variable

To summarize the calculation for a single mean, we find

$$\mathbf{d} = \frac{t_p s_{\overline{X}}}{s_X} \tag{17.2}$$

where **d** is the difference between $\overline{X}$ and the outer limits of the interval estimate expressed in terms of the number of standard deviations of the variable. This formula is very convenient if we calculate the ordinary confidence interval first. If you prefer to go directly to this interpretation, it is a little faster to use an alternative formula:

$\mathbf{d} = \frac{t_p s_{\overline{X}}}{s_X} = \frac{t_p}{\sqrt{n}}$

formulas for d

$$\mathbf{d} = \frac{t_p}{\sqrt{n}} \tag{17.3}$$

Point of Controversy

Objectivity and Subjectivity in Inferential Statistics: Bayesian Statistics

Suppose you wake up in the middle of the night with a headache. In the dark, you find three bottles in the medicine cabinet and by touch decide that you have the correct tablet. You take it and go back to bed. Later you awake with chills, sweats, and nausea. When you turn on the lights, you discover that two of the bottles you got out are aspirin but the third is poison. What is the probability that you took poison?

We introduced probability in Chapter 10. By the model given there, the probability of your having taken poison is 1/3 or .333. But suppose you had additional information. Suppose, for example, that you believe (based on reading medical news) the probability of having these symptoms to be .80 after taking poison but only .05 after taking aspirin. In the 18th century, English clergyman and mathematician Thomas Bayes introduced a theorem to deal with conditional probabilities like these (Bayes, 1763). (Note 10.2 briefly discusses conditional probability.) According to Bayes' theorem, the probability of your having taken poison is actually .89. We will not concern ourselves here with the mathematical formula for Bayes' theorem, but the essence is that *what we know after the event* (e.g., experiencing symptoms likely to be caused by poison) *can alter initial probabilities.*

Today, there is a school of thought that is known as **Bayesian statistics**, whose followers disagree with the "classical" approach to inferential statistics (taught in this textbook). The classical approach attempts to treat statistical decisions with great objectivity.

> Science earns its reputation for objectivity by treating the perils of subjectivity with the greatest of respect. (K. C. Cole, 1985)

Bayesian statistics "support inferences without reference to either significance tests or confidence intervals" (Pruzek, 1997), and offer yet another alternative to null hypothesis significance testing (e.g., Krueger, 2001).

Bayesian statistics
entails "use of prior information and empirical data to generate posterior distributions that in turn serve as the basis for statistical inferences" (Pruzek, 1997)

To compare these two approaches, consider the following research question (a hypothetical example by Berger & Berry, 1988): Is vitamin C effective in treating the common cold? Standard statistical practice is to test the null hypothesis that vitamin C has no effect (in the population). Suppose, for example, that we match 17 pairs of subjects, all suffering from cold symptoms, and randomly select one subject in each pair (person C) to receive vitamin C and the other (person P) to receive a placebo. After 24 hours, a group of physicians judge C or P to be healthier, with no ties allowed. If the null hypothesis is true, one would expect the number of pairs in which C is better to be about equal to the number of pairs in which P is better. If vitamin C works, there should be more pairs in which C is better.

How many more? Because there are only two possible outcomes (each with a probability of 1/2), the binomial distribution gives us the probabilities of outcomes of C being healthier than P. For a two-tailed test with $\alpha = .05$, we will reject the null hypothesis if C is judged to be better in fewer than 5, or more than 12, of the pairs. (The sum of the probabilities of these outcomes is .049.) Suppose that the outcome of the vitamin C experiment is 13 pairs in which C is judged to be better than P. We would reject the null hypothesis and, following standard statistical practice, express the outcome as a p-value ($p < .05$). Classical statisticians believe this to be an objective way in which to evaluate hypotheses. Bayesians contend that not only do dichotomous tests of significance limit us to a single choice—true or false—but also that the idea of rejecting a hypothesis at some percentage level is not a logical disproof (Howson & Urbach, 1991).

Bayesians prefer to acknowledge subjectivity in statistical analysis. For them, probability is not defined in terms of relative frequency but rather in degree of belief. In addition, there is no alternative hypothesis. The goal of Bayesian statistics is to state a "final" (or "posterior") probability for the hypothesis under test. The "initial" (or "prior") probability of the hypothesis being true, as well as the upper limits of this probability, are subjectively chosen *by the person interpreting the data*. In the vitamin C experiment, for example, we might believe the probability that vitamin C is effective to be .9, whereas someone else might believe it to be only .1. Some might argue that to be objective we should start with a value of .5. Bayesian analysis treats all initial hypotheses (designated H) equally and lets the data determine the final probability. This is accomplished by using the Bayes factor—the odds against the hypothesis being true provided by the data. (If you wish to read a good introduction to Bayesian analysis, see Gelman, 1998, or Pruzek, 1997). If one assumes an initial probability of .5 for H (which says that vitamin C has no effect) and chooses an upper limit of .6 (that vitamin C is effective), a final outcome of 13 pairs (out of 17) in which C is judged to be better results in a final probability of .41 for H. If one initially chooses an upward limit of .9 (a much stronger belief that vitamin C is effective), the final probability of H is .21. Essentially, Bayesian methods entail "use of prior information and empirical data to generate posterior distributions that in turn serve as the basis for statistical inferences" (Pruzek, 1997). Note that even in the case where the initial upward probability is .9, the final probability of H is still much higher than the p-values reported in the standard approach (for significant results). Thus, Bayesians contend that it is wrong to interpret small p-values as strong evidence against H_0.

The techniques discussed here (as well as throughout the text) emphasize the analysis of single studies, but, in the long run, science is concerned with series of studies. Replication of results is of foremost importance in science. Statisticians who employ the classical approach say that replication results in objective data and that there is no need for prior probabilities (e.g., Nelder, 1988). Bayesians counter that corroborating a hypothesis does not strengthen it, but only strengthens our subjective belief in it (Howson & Urbach, 1991).

17.7 Confidence Intervals for $\mu_X - \mu_Y$

In Section 17.2 you learned how to construct a confidence interval for μ_X. Just as we can make a confidence interval of μ_X, we can also estimate the true difference between μ_X and μ_Y. The logic and general nature of the procedure are the same as for the single mean.

$(\overline{X} - \overline{Y}) \pm t_p s_{\overline{X}-\overline{Y}}$
formula for constructing a confidence interval for $\mu_X - \mu_Y$

RULE FOR CONSTRUCTING A CONFIDENCE INTERVAL
FOR $\mu_X - \mu_Y$ WHEN σ_X AND σ_Y ARE UNKNOWN

$$(\overline{X} - \overline{Y}) \pm t_p s_{\overline{X}-\overline{Y}} \tag{17.4}$$

where: $(\overline{X} - \overline{Y})$ = the difference between the means of two random samples
$s_{\overline{X}-\overline{Y}}$ = the estimate of the standard error of the difference between two means
t_p = the magnitude of t for which the probability is p of obtaining a value so deviant or more so (in either direction)
$p = (1 - C)$, where C is the confidence coefficient.

If the difference $(\overline{X} - \overline{Y})$ is not farther away from $(\mu_X - \mu_Y)$ than $t_p s_{\overline{X}-\overline{Y}}$, the claim will be correct. Otherwise, it will not. Because 95% of the obtained differences (under random sampling) between the two sample means will be within this range, 95% of the claims made by following this procedure will be correct. Accordingly, if we draw a pair of samples at random, the probability is .95 that an interval estimate constructed by this procedure, $(\overline{X} - \overline{Y}) \pm t_p s_{\overline{X}-\overline{Y}}$, will include the true value of $\mu_X - \mu_Y$.

Problem

A clinician claims to have a new and better therapy to treat paranoid schizophrenia. We randomly divide 30 institutionalized patients into two samples of 15 each; one will receive the new therapy and the other the standard therapy. Because of physical illness, 2 patients in the new therapy group do not complete the study. We will refer to the new therapy group as group *X* and the standard therapy group as group *Y*. We wish to calculate the 99% confidence interval for the following data (high scores indicate greater paranoia):

NEW THERAPY (*X*)	STANDARD THERAPY (*Y*)
$\overline{X} = 87.0$	$\overline{Y} = 92.0$
$SS_X = 1200$	$SS_Y = 2016$
$n_X = 13$	$n_Y = 15$

Solution

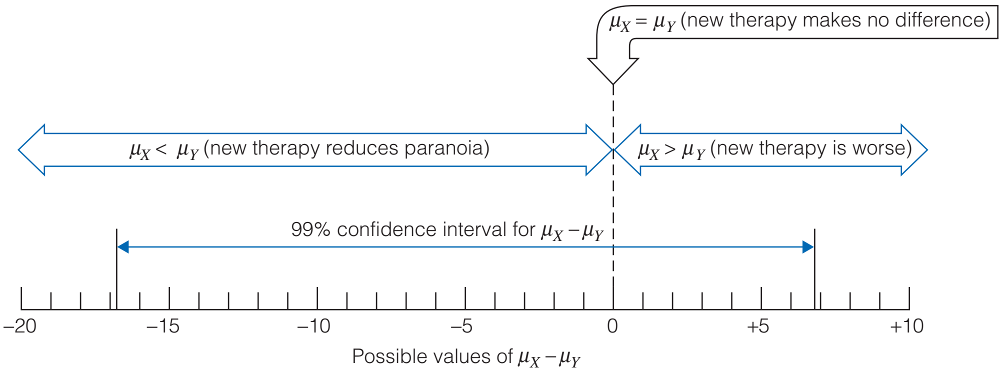
For independent samples, $df = (n_X - 1) + (n_Y - 1)$, and we must determine t_p accordingly. In the present problem, $df = 26$, and the critical value of t is $t_{.01} = 2.779$. The standard error of the difference between the means is given by Formula 14.4 (or Formula 14.5 if $n_X = n_Y$):

$$s_{\overline{X}-\overline{Y}} = \sqrt{\frac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}$$

We find that $s_{\overline{X}-\overline{Y}} = 4.21$. The 99% confidence interval is therefore constructed as follows:

$$(\overline{X} - \overline{Y}) \pm t_{.01} s_{\overline{X}-\overline{Y}}$$

LOWER LIMIT	UPPER LIMIT
$= (87.0 - 92.0) - (2.779)(4.21)$	$= (87.0 - 92.0) + (2.779)(4.21)$
$= -5.0 - 11.70$	$= -5.0 + 11.70$
$= -16.70$	$= +6.70$



$\mu_X = \mu_Y$ (new therapy makes no difference)
$\mu_X < \mu_Y$ (new therapy reduces paranoia)
$\mu_X > \mu_Y$ (new therapy is worse)
99% confidence interval for $\mu_X - \mu_Y$
–20
–15
–10
–5
0
+5
+10
Possible values of $\mu_X - \mu_Y$

FIGURE 17.4 The 99% confidence interval for the true difference in paranoia ratings in the therapy study.

Formally, this is expressed as

$$C[-16.70 \leq (\mu_X - \mu_Y) \leq +6.70] = .99$$

A negative value for $\mu_X - \mu_Y$ indicates μ_Y is larger than μ_X, whereas a positive value indicates μ_X is larger than μ_Y. Consequently, we can be 99% confident that the true difference between the mean paranoia ratings will fall anywhere between 16.70 points in favor of the new therapy (low scores are favorable here) to 6.70 points in favor of the standard therapy. This is shown in Figure 17.4. Note that a difference of 0, or no difference at all, falls within the interval.

Recall from Section 17.6 that we can express confidence intervals in terms of number of standard deviations of the variable rather than in terms of raw-score units. To reexpress the difference in standard deviations, we calculate

$$\mathbf{d} = \frac{t_p s_{\overline{X}-\overline{Y}}}{s_{\text{av}}} \qquad (17.5)$$

d
the difference between $(\overline{X} - \overline{Y})$ and the outer limits of the interval estimate, expressed in terms of the number of standard deviations of the variable

$\mathbf{d} = \dfrac{t_p s_{\overline{X}-\overline{Y}}}{s_{\text{av}}}$
formula for **d**

s_{av}
symbol for the average of s_X and s_Y

where **d** is the difference between $(\overline{X} - \overline{Y})$ and the outer limits of the interval estimate, expressed in terms of the number of standard deviations of the variable, and s_{av} is the average of s_X and s_Y (which is reasonably satisfactory if n_X and n_Y are approximately the same size). Applying Formula 17.5 to the confidence interval in our example, we obtain

$$s_X = \sqrt{\frac{SS_X}{n_X - 1}} = \sqrt{\frac{1200}{12}} = 10$$

$$s_Y = \sqrt{\frac{SS_Y}{n_Y - 1}} = \sqrt{\frac{2016}{14}} = 12$$

$$\mathbf{d} = \frac{(2.779)(4.21)}{(10.0 + 12.0)/2} = 1.06$$

This informs us that the true difference between the two means is within 1.06 standard deviations of the obtained difference according to our 99% confidence interval.

If we desire, we can restate the difference relative to a difference of zero rather than to the obtained difference between the two sample means. To do this, we simply divide the upper and lower raw-score limits of the confidence interval by s_{av}. In the present example, the lower limit becomes $-16.70/11.0 = -1.52$, and the upper limit becomes $+6.70/11.0 = .61$. This states, with 99% confidence, that the difference between the two population means is such that μ_X could be lower than μ_Y by 1.52 standard deviations, that μ_X could be higher than μ_Y by .61 standard deviation, or that the difference falls between these limits. When confidence limits are expressed this way, we need to keep in mind that the width of the limits must be considered in the light of the value of the confidence coefficient employed, just as when the limits are expressed in score points.

When our samples are *dependent*, the procedures are the same except for the calculation of $s_{\overline{X}-\overline{Y}}$. For dependent samples, it is given by Formula 15.2:

$$s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2 - 2r_{XY}s_{\overline{X}}s_{\overline{Y}}}$$

For our example about comparing methods of teaching arithmetic (Chapter 15), we calculated $s_{\overline{X}-\overline{Y}}$ to be 1.76, $\overline{X} = 69.73$, $\overline{Y} = 65.07$, and $t_{.05} = \pm 2.145$ when $df = 14$. Therefore, to construct the 95% confidence interval:

$$(\overline{X} - \overline{Y}) \pm t_{.05}\ s_{\overline{X}-\overline{Y}}$$

LOWER LIMIT	UPPER LIMIT
$= (69.73 - 65.07) - (2.145)(1.76)$	$= (69.73 - 65.07) + (2.145)(1.76)$
$= 4.66 - 3.78$	$= 4.66 + 3.78$
$= .88$	$= 8.44$

$$C[0.88 \leq (\mu_X - \mu_Y) \leq 8.44] = .95$$

If we want, we may also construct confidence intervals from difference scores:

$\overline{D} \pm t_p s_{\overline{D}}$ formula for constructing a confidence interval for μ_D

$$\overline{D} + t_p\ s_{\overline{D}}$$

From Table 15.1, we calculate $s_{\overline{D}}$ to be $\sqrt{3.10} = 1.76$. Therefore,

$$4.67 \pm (2.145)(1.76)$$

and we arrive at the same answer.

Remember, *the confidence interval contains all values of* H_0 *that would be retained had they been tested using* $\alpha = 1 - C$ (for a *nondirectional* alternative hypothesis). Consider, for example, a problem in which $\overline{X} - \overline{Y} = +12$, $s_{\overline{X}-\overline{Y}} = 5$, and $n = 25$ for both groups. When the null hypothesis of zero difference is tested, $t = 12/5 = +2.4$; the difference is significant at the .05 level ($t_{crit} = \pm 2.02$ for 48 *df*). Calculating a confidence interval from exactly the same data produces $C[+1.9 \leq (\mu_X - \mu_Y) \leq +22.1] = .95$. This confidence interval says that the value of μ_X exceeds that of μ_Y by 1.9 to 22.1 points. Note that the interval does *not* include the possibility that the difference could be zero, the same message conveyed by the test of the hypothesis. In addition, the confidence interval informs us what the true difference *could* be.

Consider a second example: If a confidence interval reads $C[-16.70 \leq (\mu_X - \mu_Y) \leq +6.70] = .99$, such as we found for the example at the start of this section, zero is among the possible values that the difference might take. Had we tested the null hypothesis of no difference at $\alpha = .01$, it would have produced a t smaller than the value needed to reject the hypothesis.

Here is a third confidence interval: $C[+1 \leq (\mu_X - \mu_Y) \leq +4] = .95$. It shows that the true difference is not likely to be zero; instead, μ_X is probably higher than μ_Y by 1 to 4 points. The difference is large enough that chance variation cannot account for it, but if the standard deviations of X and Y are about 15, it is nevertheless a small difference.

17.8 Sample Size Required for Confidence Intervals of μ_X and $\mu_X - \mu_Y$

Suppose we wish to estimate μ_X such that it will not be farther away from $\overline{X}$ than a given amount, with confidence coefficient equal to .95. If we can estimate σ_X before we begin, it is possible to estimate the size of the sample that will be required to produce this result. Required sample size is given by

$n = \left(\frac{\sigma_X z_p}{w}\right)^2$
formula for calculating sample size required for a confidence interval of μ_X of given width

w
symbol for the maximum distance (in score points) desired between $\overline{X}$ and the limits of the estimate of μ_X

SAMPLE SIZE REQUIRED FOR A CONFIDENCE INTERVAL OF μ_X OF GIVEN WIDTH

$$n = \left(\frac{\sigma_X z_p}{w}\right)^2 \qquad (17.6)$$

where **w** is the maximum distance (in score points) desired between $\overline{X}$ and the limits of the estimate of μ_X. If σ_X is estimated to be 16 and we aim to estimate μ_X with 95% confidence that it will not be farther away from $\overline{X}$ than 4 points, the calculation will be

$$n = \left[\frac{(16)(1.96)}{4}\right]^2 = 61$$

We can often make an approximate estimate of σ_X by referring to previous studies that used the same variable. Even if we are using a new variable, it may be possible to make some kind of guess.

When construction of the interval depends on estimating σ_X from a sample, the estimated sample size tends to be an underestimate, especially if the required sample size is small. In this case, it is best to use a slightly larger sample than that given by Formula 17.6. In fact, whenever it is important to be conservative, take more cases than Formula 17.6 suggests.

We can also estimate $\mu_X - \mu_Y$ with a given degree of confidence that it will not be farther away from $(\overline{X} - \overline{Y})$ than a given amount. For independent samples, when samples are of (nearly) equal size and it is reasonable to believe that $\sigma_X = \sigma_Y$ (or at least approximately so), we find the required size of *each* of the two samples by calculating sample size according to Formula 17.6 *and multiplying that result by 2*:

$n = \left(\frac{\sigma z_p}{w}\right)^2(2)$
formula for calculating sample size required for a confidence interval of $(\mu_X - \mu_Y)$ of given width

SAMPLE SIZE REQUIRED FOR A CONFIDENCE INTERVAL OF $(\mu_X - \mu_Y)$ OF GIVEN WIDTH

$$n = \left(\frac{\sigma z_p}{w}\right)^2(2) \qquad (17.7)$$

where: w = maximum desired discrepancy (in score points) between $(\overline{X} - \overline{Y})$ and the limits of the estimate of $\mu_X - \mu_Y$.

TABLE 17.1 *Approximate Maximum Distance (w) Measured in σ Units, Expected According to Selected Values of Sample Size between (1) $\overline{X}$ and the Limits of the 95% Confidence Interval for Single Means and (2) $\overline{X} - \overline{Y}$ and the Limits of the 95% Confidence Interval for Two Independent Means**

	SAMPLE SIZE								
	10	*15*	*25*	*50*	*75*	*100*	*200*	*500*	*1,000*
Single means: $\lvert\overline{X} - \mu_X\rvert / s_X$	.72	.55	.41	.28	.23	.20	.14	.09	.06
Difference between two independent means: $\lvert(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)\rvert / s_{\text{av}}$	.94	.75	.57	.40	.32	.28	.20	.12	.09

*For two independent samples, n is the size of *each* of the two samples. It is assumed that $n_X = n_Y = n$ and that $\sigma_X = \sigma_Y$.

Again, in practical applications it is usually possible to make at least an approximate estimate of σ.

Suppose we hope to estimate $\mu_X - \mu_Y$ with 95% confidence that it will not be farther away from $(\overline{X} - \overline{Y})$ than 4 score points and that we can safely assume $\sigma_X = \sigma_Y = 16$. What sample size will be required? The calculation is

$$n = 2\left[\frac{(16)(1.96)}{4}\right]^2 = 123$$

Note that (within rounding error) *twice* as many cases are required for *each* sample to maintain the same distance, w, for the two-sample estimate as were required for the one-sample estimate.

If samples are dependent, we find the number of *pairs* of elements by the same procedure, except that we multiply the resulting value of n by $(1 - \rho)$. Once again, we will have to estimate ρ (the value of Pearson's correlation coefficient for the population).

Table 17.1 gives an idea of the degree of accuracy that we may achieve in interval estimation from samples of selected sizes. The entries represent w, expressed as a fraction of the standard deviation of the variable, for 95% confidence intervals. For example, the table says that if we use 100 cases in estimating $(\mu_X - \mu_Y)$, the 95% confidence interval will be $(\overline{X} - \overline{Y}) \pm .28s_{\text{av}}$. Treat the values in the tables as approximations, particularly for the lowest three values of n.

17.9 Confidence Intervals for ρ

If we want to know within what limits the population coefficient is found (rather than testing the hypothesis that ρ is some specific value), we can construct a confidence interval by translating the sample r to z' and using the following rule:

$z' = z_p\sigma_{z'}$
formula for constructing a confidence interval for ρ

RULE FOR CONSTRUCTING A CONFIDENCE INTERVAL FOR ρ

$$z' = z_p\sigma_{z'} \tag{17.8}$$

where:
z' = the value of Fisher's z' corresponding to the sample r
z_p = the magnitude of normally distributed z for which the probability is p of obtaining a value as deviant or more so (in either direction)
$p = (1 - C)$, where C is the confidence coefficient
$\sigma_{z'}$ = the standard error of Fisher's z'.

The formula for the standard error of z' is

STANDARD ERROR OF FISHER'S z'

$$\sigma_{z'} = \frac{1}{\sqrt{n-3}} \qquad (17.9)$$

$\sigma_{z'}$ symbol for the standard error of Fisher's z'; $=1/\sqrt{n-3}$

Formula 17.8 results in a lower limit and an upper limit, both expressed in terms of z'. We must then convert these to values of r using Table H in Appendix D.

We illustrate the procedure with the same data we used in Section 16.2. Recall that the sample r was $+.30$, based on 25 pairs of observations. What are the limits that we can be 95% confident include ρ, the population value of the correlation coefficient? From Table H, we find that $z' = +.31$ corresponds to $r = +.30$. The standard error is $\sigma_{z'} = 1/\sqrt{25-3} = .21$, and 95% of the values in a normal distribution are within ± 1.96 standard errors (z) of the mean. Thus, according to Formula 17.8, the limits of the 95% confidence interval are

$$\begin{array}{lcc} & \text{Lower Limit} & \text{Upper Limit} \\ +.31 \pm (1.96)(.21) = & -.10 \text{ to} & +.72 \end{array}$$

Using Table H, we translate these values of z' back to values of r. The resulting limits are

$$r_{LL} = -.10 \qquad r_{UL} = +.62$$

Formally, we express the confidence interval as

$$C(-.10 \le \rho \le +.62) = .95$$

Note that these limits are not equidistant from $r = +.30$. The upper limit is closer to $r = +.30$ because the sampling distribution of r is skewed (see Section 16.1).

The use of confidence intervals shows us why correlation coefficients based on small samples are usually not worth much. Note, for instance, that when n is 10 ($df = 8$) and $r = 0$, the 95% confidence limits for r are $\pm .63$. Suppose that, unknown to us, $\rho = +.60$ and that our sample r happens to have exactly the same value. Nevertheless, if we are testing the hypothesis that $\rho = 0$, we would find no cause to reject it. Even when $n = 32$, the 95% limits are $\pm .35$. It should be obvious that we need a good-sized sample to determine the extent of relationship with any degree of accuracy. At the other extreme, if samples are very large, a relatively small coefficient may lead to the mistaken conclusion that there is a strong relationship ("highly significant") between two variables when the hypothesis of no correlation is tested. Recall, for example, the study that found a "definite" link between children's height and IQ (see Section 13.2). Interval estimation can help researchers avoid the traps inherent in using either small or large samples. *It would be good practice to use confidence intervals of ρ much more often than is commonly done in research.*

17.10 Where Are We in Statistical Reform?

Change often comes slowly, and this is certainly true for statistical reform in inferential statistics. Despite the many calls for less emphasis on null hypothesis significance testing (e.g., Cohen, 1994; Cumming, 2008; Hunter, 1997; Kehle et al., 2007; Loftus, 1996; Schmidt, 1996), nearly all studies published in psychology journals continue to use hypothesis testing (e.g., Cumming et al., 2007; Kehle et al., 2007). However, more published articles are now including figures with error bars (standard errors or confidence intervals) and many journals now require that authors include effect sizes. Our advice: do not wait for journal editors to require effect sizes and confidence intervals even if you use null hypothesis significance testing. Inclusion of these measures is a good statistical practice.

17.11 Summary

Prior to this chapter, we developed models for testing hypotheses about a specific value of the population mean (Chapter 12), the difference between two population means (Chapters 14 and 15), and the population correlation coefficient (Chapter 16). In establishing confidence intervals, no value is specified in advance. Instead, we draw a sample and calculate an interval of values within which it appears reasonable that the population parameter lies (interval estimation). Any problem for which hypothesis testing is useful can also be worked by confidence intervals, but the reverse is not true.

In a normal distribution, 95% of the scores are within 1.96 standard deviations of the mean. When we draw random samples of a particular size from such a distribution, 95% of the sample means are within $\pm 1.96\sigma_{\overline{X}}$ of μ_X. Therefore, it is also the case that for 95% of sample means, μ_X is no farther away than $1.96\sigma_{\overline{X}}$. In making estimates from sample data, we must, as usual, use $s_{\overline{X}}$ in place of $\sigma_{\overline{X}}$ and the t distribution in place of the normal curve. The interval in which we are reasonably confident of finding μ_X is thus found by the formula $\overline{X} \pm t_p s_{\overline{X}}$, where t_p is the magnitude of t for which the probability is p of obtaining a value as deviant or more so in either direction. The limits of the confidence interval will vary from sample to sample, but when $p = .05$, 95% of the interval estimates will contain μ_X. To construct a confidence interval for the difference between the means of two populations, the formula is $(\overline{X} - \overline{Y}) \pm t_p s_{\overline{X}-\overline{Y}}$.

If we are not familiar enough with the variable under study to understand whether the estimated interval is large or small, we may express the interval limits in terms of the standard deviation of the variable. If we want to estimate μ_X such that it will not be farther away from $\overline{X}$ than a given amount, we can calculate the sample size necessary if we can estimate σ_X in advance of collecting our data.

Confidence intervals of ρ are a better method of presenting correlational coefficients because they directly show the influence of sampling variation. To compute confidence intervals, we must first convert r to another statistic, z', because the sampling distribution of r is not normally distributed.

Key Terms, Concepts, and Symbols

dichotomous decision making (p. 282)
p-value (p. 282)
point estimates (p. 283)
interval estimates (p. 283)
t_p (p. 284)
confidence interval (p. 285)
confidence coefficient (p. 285)
C (p. 285)
d (p. 289)
Bayesian statistics (p. 289)
w (p. 294)
$\sigma_{z'}$ (p. 296)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Given: $\overline{X} = 63$, $s_X = 12$, $n = 100$. Construct a confidence interval of μ_X according to: (a) $C = .95$ and (b) $C = .99$.

2. Repeat parts (a) and (b) of Problem 1, but with the condition that n is four times as large. Compare the size of the intervals obtained in the two problems. State your conclusion.

3. Given: $\overline{X} = 117$, $s_X = 36$, $n = 81$. Construct a confidence interval of μ_X according to (a) $C = .95$ and (b) $C = .99$.

4. Using the data of Problem 1, interpret a confidence interval estimate of μ_X in terms of **d**, using $C = .95$. Does it appear that the estimate has considerable precision or not? Explain.

5. Repeat Problem 4, but use $n = 25$.

6. Given: $\overline{X} = 2.90$, $s_X = .30$, $n = 16$. (a) Construct a confidence interval of μ_X using $C = .95$. (b) What is your feeling about the precision of this estimate? (c) Calculate **d**. (d) What is your feeling now about the precision of this estimate? (e) Comment on the appropriateness of applying the techniques of this chapter to this problem.

7. Given: $\overline{X} = 165$, $s_X = 24$, $n_X = 36$, $\overline{Y} = 175$, $s_Y = 21$, $n_Y = 49$; assume the samples are independent. Construct a confidence interval of $\mu_X - \mu_Y$ according to (a) $C = .95$ and (b) $C = .99$.

8. Given: $\overline{X} = 97$, $s_X = 16$, $n_X = 64$, $\overline{Y} = 90$, $s_Y = 18$, $n_Y = 81$; assume the samples are independent. Construct a confidence interval of $\mu_X - \mu_Y$ according to (a) $C = .95$ and (b) $C = .99$.

9. Using the data of Problem 7, interpret a confidence interval of $\mu_X - \mu_Y$ in terms of **d**, using $C = .95$. Does it appear that the estimate has considerable precision? Explain.

10. Construct the 95% confidence interval for the data of Problem 7, expressing the upper and lower limits in terms of the number of standard deviations relative to a difference of zero between μ_X and μ_Y.

11. Repeat Problem 9 using the data of Problem 8.

12. Given: $C(122 \leq \mu_X \leq 128) = .95$. (a) If H_0: $\mu_X = 127$ is tested at $\alpha = .05$ using a nondirectional alternative, what is the outcome? (b) If H_0: $\mu_X = 118$ is tested at $\alpha = .05$, what is the outcome? (c) Explain your answers to (a) and (b).

13. Given: $C[-8 \leq (\mu_X - \mu_Y) \leq -2] = .95$. (a) What is $\overline{X} - \overline{Y}$? (b) If H_0: $\mu_X - \mu_Y = 0$ is tested at $\alpha = .05$ using a nondirectional alternative, what is the outcome?

14. Given: $C[-2 \leq (\mu_X - \mu_Y) \leq +12] = .95$. (a) What is $\overline{X} - \overline{Y}$? (b) If H_0: $\mu_X - \mu_Y = 0$ is tested at $\alpha = .05$ using a nondirectional alternative, what is the outcome?

15. Given: $\overline{X} = 83$, $s_X = 15$, $n = 625$. (a) Obtain z corresponding to a test that $\mu_X = 80$. (b) What, in "loose English," does this value for z tell us? (c) Construct the 95% confidence interval of μ_X. (d) What does it tell us? (e) Compare the information obtained by the two approaches.

16. Construct the 95% confidence interval for the data in Problem 7, Chapter 14.

17. Construct the 99% confidence interval for the data in Problem 9, Chapter 14.

18. Construct the 99% confidence interval for the data in Problem 10, Chapter 14.

19. Construct the 95% confidence interval for the data in Problem 11, Chapter 14.

20. Construct the 95% confidence interval for the data in Problem 4, Chapter 15.

21. Construct the 95% confidence interval for the data in Problem 5, Chapter 15.

22. Construct the 99% confidence interval for the data in Problem 6, Chapter 15.

23. Construct the 99% confidence interval for the data in Problem 7, Chapter 15.

24. A psychologist wishes to estimate mean reaction time such that he can be 95% confident that μ_X is not farther than 3 milliseconds (msec) away from his sample mean. The psychologist estimates that the standard deviation of reaction times is 15 msec. What size sample should he draw?

25. Repeat Problem 25, but with the condition that the confidence coefficient should be .99.

26. With reference to Problem 25, our psychologist now wishes to change the nature of the reaction task and to compare mean reaction time on the original task with that on the altered task. He assumes that the standard deviation will be the same on both tasks and wants to estimate at the 95% level of confidence the difference between the two population means. The psychologist wants this difference to be within 5 msec of the difference between the sample means. (a) If the psychologist selects two random samples, what should be the size of each? (b) If he uses subjects matched on a pretest of reaction time and ρ is estimated to be +.40 between pairs of measurements, what should be the size of each sample?

27. For $r = +.25$, construct the 95% confidence intervals using Formula 17.8 when (a) $n = 10$, (b) $n = 20$, (c) $n = 40$, and (d) $n = 80$.

28. Repeat Problem 28 for $r = +.65$.

CHAPTER 18

Testing for Differences among Three or More Groups: One-Way Analysis of Variance (and Some Alternatives)

When you have finished studying this chapter, you should be able to:

- Understand that using multiple *t* tests to make comparisons among three or more groups increases the likelihood of making a Type I error;
- Explain how analysis of variance uses estimates of the variance to test hypotheses about differences among population means;
- Calculate SS_{total}, SS_{w}, and SS_{bet} as well as their respective degrees of freedom;
- Understand that s^2_{w} is always a measure of inherent variation, whereas s^2_{bet} is a measure of inherent variation plus any treatment effect;
- Construct a summary table and calculate the *F* ratio;
- Use IBM SPSS to calculate the *F* ratio;
- Understand the assumptions underlying ANOVA;
- Calculate effect size;
- Make *post hoc* comparisons, when *F* is statistically significant, to determine which groups differ;
- Know when and how to use planned comparisons as an alternative to ANOVA; and
- Calculate analysis of variance for repeated measures.

In Chapters 14 and 15, you learned how to use the t distribution to test the hypothesis of no difference between *two* population means. Suppose, however, that we wish to know about the relative effect of three or more different "treatments." Are there differences in effectiveness among three methods of conducting psychotherapy? Among seven experimental drugs being tested for use with Alzheimer's patients, are there differences in their effects on mental ability?

In both of these examples, we could use the t test to make comparisons among each possible combination of two means. However, this method is not acceptable for several reasons. Consider the example involving seven experimental drugs to treat Alzheimer's patients. First, we would have to conduct 21 t tests if each drug is to be compared with each of the others. Second, with so many tests, there is an increased likelihood of committing at least one Type I error—that is, of obtaining a "significant difference" when no true difference exists. Suppose, for example, that none of the drugs is effective. If we set the level of significance at .05 for each individual test, the probability of a Type I error is .05 for each. However, taken as a group, the probability that at least one from among the 21 t tests will prove to be a false positive is greater than .05, and it increases with additional tests. In fact, if all 21 tests are *independent*, the probability of at least one false positive is .66 [calculated by $p = 1 - (.95)^{21}$]. When tests are not independent, it is impossible to calculate the exact probability of committing at least one Type I error, but it is quite high. Thus, if we find one outcome to be "significant at the .05 level" among 21 tests, we should not consider that this result is mined gold.

Type I error rejection of a true null hypothesis

analysis of variance a statistical procedure used to compare simultaneously the means of two or more populations

ANOVA abbreviation for "analysis of variance"

one-way analysis of variance an analysis of variance with only one independent variable; a single-factor design

Third, in any one comparison, we use only the information provided by those two groups. When the 21 tests are completed, there are 21 bits of information, but not a single direct answer as to whether, taken as a whole, there is evidence of differences in effect among the seven drugs.

In the 1920s, Sir Ronald Fisher developed a better answer to such problems. The technique is called **analysis of variance**, abbreviated **ANOVA**. The technique allows us to compare simultaneously several means with the level of significance specified by the investigator. In this chapter, you will learn how to use the simplest form, called *one-way analysis of variance*. We introduce a more complex form in the following chapter.

Before you continue reading this chapter, you should review some topics covered in previous chapters. Particularly important topics include:

1. The concept of variance (Section 5.3).
2. Calculation of the sum of squares, $\Sigma(X - \overline{X})^2$ (Section 5.3).
3. s^2 as an estimate of σ^2 (Section 12.9).
4. Degrees of freedom in a variance estimate (Section 12.12).
5. Homogeneity of variance (Section 14.4).
6. Type I error (Section 13.3).
7. Assumptions associated with inference about the difference between two means (Section 14.11).

18.1 The Null Hypothesis

treatment conditions different values (**levels**) of the independent variable

factor a categorical independent variable

In one-way analysis of variance, there may be two or more **treatment conditions**, often referred to as different **levels** of the categorical independent variable or **factor**. We will use k to represent the number of treatment conditions. The k treatments may be identified by letters, such as A, B, and C, and the population means as μ_A, μ_B, and μ_C. If the different treatments have no effect on the variable under observation, then we expect these population means to be equal. We test the null hypothesis:

$$H_0: \mu_A = \mu_B = \mu_C = \ldots = \mu_k$$

The null hypothesis in ANOVA is often referred to as an *omnibus hypothesis* (i.e., covering many situations at once), and ANOVA itself as an *omnibus test*. The alternative hypothesis, often stated simply as "not H_0," is that the population means are different in *some* way. For example, two or more group means may be alike and the remainder different, all may be different, and so on. Recall that when we test the hypothesis of no difference between *two* means, a distinction was made between directional and nondirectional alternative hypotheses. Such a distinction no longer makes sense when the number of means exceed two.

k
symbol for the number of treatment conditions

18.2 The Basis of One-Way Analysis of Variance: Variation within and between Groups

We will first develop one-way analysis of variance for independent groups. ANOVA for repeated measures is introduced in Section 18.18.

Suppose that we have selected three samples of 10 cases each at random and have given them three different treatments. If there is *no* **treatment effect**, the null hypothesis that $\mu_A = \mu_B = \mu_C$ is true, and the distribution of scores in the three samples might appear as shown in Figure 18.1. If there *is* a treatment effect, H_0 is false, and the three samples might appear as shown in Figure 18.2, where μ_A, μ_B, and μ_C have different values. Let's compare Figures 18.1 and 18.2. For the scores in either figure, we may take the total variation among all the scores and divide it (statisticians say "partition" it) into two parts (see Figure 18.3) (we use the word *variation* here and throughout the chapter in a general sense to mean only that scores vary):

null hypothesis for ANOVA
H_0: $\mu_A = \mu_B = \mu_C \ldots = \mu_k$

treatment effect
differences among groups caused by the difference in treatment conditions

1. *Within-groups variation.* Within each sample, the individual scores vary around the sample mean. We will call this **within-groups variation**. Remember what you have learned about the influence of random sampling variation. Within-groups variation is a direct reflection of the *inherent variation* (variation due to chance) among individuals given the same treatment. We can present exactly the same stimulus to everyone in a group and still observe variation in reaction times. It's a fact of life that even under identical conditions, individuals vary in performance.

Notice that the amount of within-groups variation in Groups *A, B*, and *C* is about the same whether the three population means are identical (Figure 18.1) or different (Figure 18.2). *Within-groups variation reflects only inherent variation. It does not reflect differences caused by differential treatment.* The reason is that within a particular treatment group, each subject gets the *same* treatment. Although we have referred to within-groups variation as inherent variation, you should be aware that some statisticians call it *error*.

within-groups variation
variation of scores around the mean in a single treatment condition (here, scores can vary only because of inherent variation—chance)

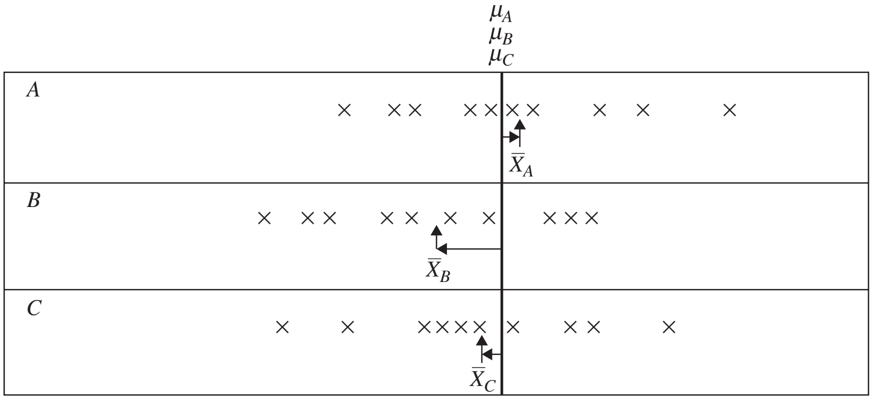
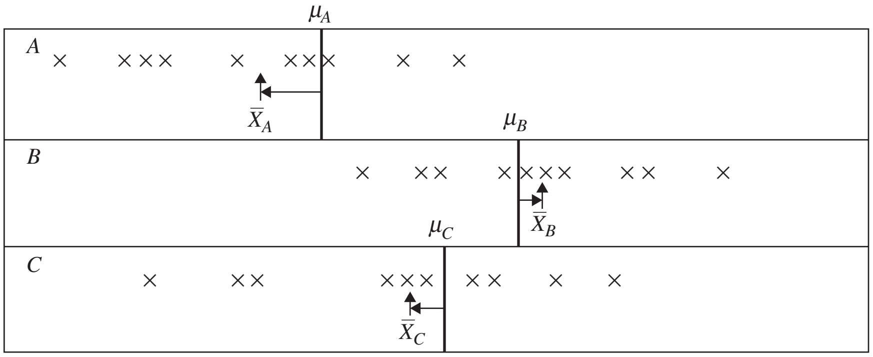
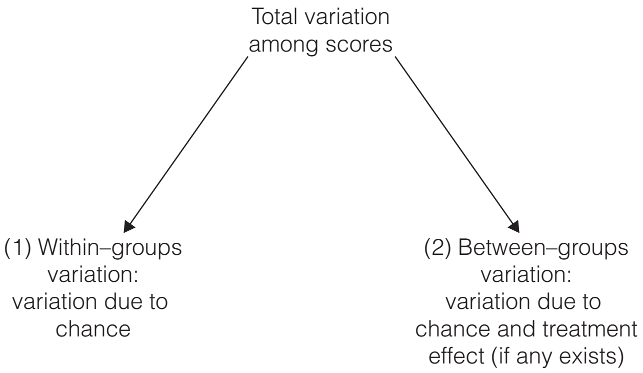


FIGURE 18.1 Distribution of 10 scores in each of three groups with no treatment effect—H_0: $\mu_A = \mu_B = \mu_C$ is true.



μ_A
A
$\overline{X}_A$
μ_B
B
$\overline{X}_B$
μ_C
C
$\overline{X}_C$

FIGURE 18.2 Distribution of 10 scores in each of three groups with treatment effect present—H_0: $\mu_A = \mu_B = \mu_C$ is false.



Total variation among scores
(1) Within–groups variation: variation due to chance
(2) Between–groups variation: variation due to chance and treatment effect (if any exists)

FIGURE 18.3 The total variation among scores (see Figures 18.1 and 18.2) may be divided into two parts: (1) variation due to chance factors (within-groups variation), and (2) variation due to chance factors plus treatment effect (between-groups variation).

between-groups variation variation among the means of the different treatment conditions (here, means can vary because of inherent variation and treatment effect, if any exists)

2. *Between-groups variation.*[1] Look again at Figure 18.1. The means of the samples vary among themselves. Of course, we do not expect $\overline{X}_A$, $\overline{X}_B$, and $\overline{X}_C$ all to have the same value even when $\mu_A = \mu_B = \mu_C$; this, too, is in accord with what you have learned about random sampling. It is important to realize that **between-groups variation** is also a reflection of the inherent (chance) variation among individuals. But what if there is a treatment effect? In that case, the sample means will also differ from one another. Thus, between-groups variation is due to both inherent variation plus treatment effect (if any exists), as is the case in Figure 18.2. Let us now develop this further.

18.3 Partition of the Sums of Squares

grand mean the mean of all scores in all the treatment conditions

When we compare two or more populations, the mean of all the scores is called the **grand mean**. We obtain it by adding all the scores in all the populations and dividing by the total number of scores. Any single score (X) in any of the populations may be thought of as the sum of three components:

$$X = \textit{grand mean} + \textit{treatment effect} + \textit{inherent variation}$$

---

[1] With more than two groups, proper English is "among groups," but "between groups" is the standard terminology in statistics.

inherent variation variation due to chance factors

$\bar{\bar{X}}$ symbol for the grand mean

Treatment effect is the amount by which a population mean deviates from the grand mean. Inherent variation (the "error effect") is everything that is left over ($X -$ *grand mean* $-$ *treatment effect*). Because all the subjects in a particular group are treated the same, **inherent variation** is the variation due to chance factors. For a score in any one group, it is $(X - \mu_X)$. When we draw samples from the populations, our sample estimates for the grand mean ($\bar{\bar{X}}$), treatment effect $(\bar{X} - \bar{\bar{X}})$, and inherent variation may be shown as:

PARTITION OF A SINGLE SCORE X

$$X = \bar{\bar{X}} + (\bar{X} - \bar{\bar{X}}) + (X - \bar{X}) \qquad (18.1)$$

Any score in any group may be expressed as a deviation from the grand mean: $(X - \bar{\bar{X}})$. In Formula 18.1, simply move the grand mean to the left of the equals sign. The resulting formula shows that how much a score deviates from the grand mean can be broken into two parts: (1) how much the score deviates from the mean of its own sample group, $(X - \bar{X})$, and (2) how much the mean of its sample group deviates from the grand mean, $(\bar{X} - \bar{\bar{X}})$. We can represent this as follows:

PARTITION OF $X - \bar{\bar{X}}$

$$X - \bar{\bar{X}} = (X - \bar{X}) + (\bar{X} - \bar{\bar{X}}) \qquad (18.2)$$

Consider, for example, three sets of scores: 1, 2, 3 ($\bar{X}_A = 2$); 4, 5, 6 ($\bar{X}_B = 5$); and 7, 8, 9 ($\bar{X}_C = 8$), where $\bar{\bar{X}} = 5$. For the score of 7:

$$\begin{aligned} 7 - 5 &= (7 - 8) + (8 - 5) \\ +2 &= -1 + 3 \\ +2 &= +2 \end{aligned}$$

We can use Formula 18.2 to develop (1) a measure of variation that reflects only inherent variation (within-groups variation) and (2) a measure that reflects inherent variation plus treatment effect, if present (between-groups variation).

SS_{total} sum of the squared deviations from the grand mean, $\sum^{\text{all scores}} (X - \bar{\bar{X}})^2$

Recall from Section 5.3 that the numerator of the variance is the sum of the squared deviations from the mean, commonly called the *sum of squares* and abbreviated SS. To calculate **SS_{total}**, the sum of the squared deviations from the grand mean, we must first square each of the deviations from the grand mean and then sum these squares. Symbolically, this is

$$SS_{\text{total}} = \sum^{\substack{\text{all}\\ \text{scores}}} (X - \bar{\bar{X}})^2$$

From Formula 18.2 we can derive another index of SS_{total} by squaring everything to the right of the equals sign in that formula. The result is:

$$\sum^{\substack{\text{all}\\ \text{scores}}} (X - \bar{\bar{X}})^2 = \sum^{\substack{\text{all}\\ \text{scores}}} (X - \bar{X})^2 + \sum^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 \qquad (18.3a)$$

where n_i is the number of scores in the ith group, $\bar{X}_i$ is the mean of the ith group, and k is the number of groups. We give a proof of this equality in Note 18.1 at the end of the chapter.

From the total sample, we thus obtain three different sums of squares:

SS_{within} within-groups (treatments) sum of squares $\sum^{\text{all scores}} (X - \bar{X})^2$

SS_{between} between-groups (treatments) sum of squares $\sum^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$

1. $\sum^{\text{all scores}} (X - \bar{\bar{X}})^2$, the total sum of squares, called **SS_{total}**.
2. $\sum^{\text{all scores}} (X - \bar{X})^2$, the within-groups sum of squares, called **SS_{within}** (sometimes called SS_{error}).
3. $\sum^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2$, the between-groups sum of squares, called **SS_{between}**.

Thus, Formula 18.3*a* can also be expressed as follows:

PARTITION OF TOTAL SUM OF SQUARES

$$SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}} \qquad (18.3b)$$

$SS_{\text{total}} = SS_{\text{within}} + SS_{\text{between}}$ partition of the sums of squares

This splitting of SS_{total} into two parts is referred to as the **partition of the sums of squares**. Let us now examine each part of Formula 18.3b more closely (see Figure 18.4):

1. *Total sum of squares.* This is a measure of the total variation present in the data without regard to group.
2. *Within-groups sum of squares.*

$$SS_{\text{within}} = \sum^{\text{all scores}} (X - \bar{X})^2 = \sum (X_A - \bar{X}_A)^2 + \sum (X_B - \bar{X}_B)^2 + \cdots \qquad (18.4)$$

assumption of homogeneity of variance the assumption that the population variances are the same for all treatment conditions

From now on, we will abbreviate SS_{within} as SS_{w}. This is a measure of the within-groups variation of individual scores, independent of any differences among group means. It is a direct reflection of *inherent variation, free from the influence of treatment effect.* An estimate could be made by taking the sum of squares for any one group. However, on the assumption that the population variances are the same for all groups ($\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \ldots$, the **assumption of homogeneity of variance**), we may obtain a better estimate by combining information from the several groups. The variance that is assumed to be common to all the populations is σ^2 (with no subscript).

3. *Between-groups sum of squares.*

$$SS_{\text{between}} = \sum^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2 = n_A(\bar{X}_A - \bar{\bar{X}})^2 + n_B(\bar{X}_B - \bar{\bar{X}})^2 + \cdots \qquad (18.5)$$

From now on, we will abbreviate SS_{between} as SS_{bet}. This is a measure of the variation of the sample means among groups. It reflects inherent variation (free from the influence

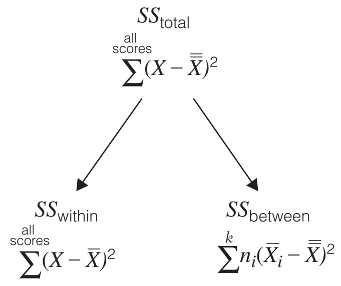

SS_{total}
$\sum^{\text{all scores}}(X-\bar{\bar{X}})^2$
SS_{within}
$\sum^{\text{all scores}}(X-\bar{X})^2$
SS_{between}
$\sum^{k} n_i(\bar{X}_i-\bar{\bar{X}})^2$

FIGURE 18.4 Partitioning of the sums of squares for a one-way analysis of variance for independent samples.

of treatment effect) *plus* any treatment effect (variation attributable to differences among population means).

18.4 Degrees of Freedom

The test of H_0 in one-way ANOVA involves comparing two separate estimates of the variance for the population: one based on the within-groups variation among individual scores about their group means, and the other on the variation among the sample means. We introduced the subject of estimating population variances in Section 12.9. Recall that to calculate the unbiased estimate of a population variance, we divide the sum of squares by the degrees of freedom associated with that sum of squares. This general relationship is summarized by the equation

$$s^2 = \frac{SS}{df}$$

Recall also that the degrees of freedom for the variance estimate based on a single sample is $n - 1$, and that the degrees of freedom for the pooled estimate of the variance used in the t test for two independent means is $(n_1 - 1) + (n_2 - 1)$.

df_{total}
$n_{\text{total}} - 1$

The degrees of freedom associated with SS_{total} is, as you might expect, $n_{\text{total}} - 1$, where n_{total} is the total number of cases. Because there are $n - 1$ degrees of freedom associated with the deviations about a single mean, the degrees of freedom for SS_{w} is $(n_A - 1) + (n_B - 1) + \cdots = n_A + n_B + \cdots + n_k - k$, or

df_{within}
$n_{\text{total}} - k$

$$df_{\text{w}} = n_{\text{total}} - k$$

where k is the number of groups. For SS_{bet}, there are only as many deviations from the grand mean as there are sample means (or number of groups). Thus,

df_{between}
$k - 1$

$$df_{\text{bet}} = k - 1$$

For *any* analysis of variance, the number of degrees of freedom associated with SS_{w} plus the number of degrees of freedom associated with SS_{bet} equals the number of degrees of freedom associated with SS_{total}:

$df_{\text{total}} = df_{\text{w}} + df_{\text{bet}}$
partition of degrees of freedom

$$df_{\text{total}} = df_{\text{w}} + df_{\text{bet}}$$

We give proof of this equality in Note 18.1 at the end of the chapter.

18.5 Variance Estimates and the *F* Ratio

within-groups variance estimate
an estimate of inherent variance only

If we now divide SS_{w} and SS_{bet} by their respective degrees of freedom, we obtain the two variance estimates required to test H_0: $\mu_A = \mu_B = \mu_C$, the **within-groups variance estimate**, s^2_{w}, and the **between-groups variance estimate**, s^2_{bet}. These variance estimates, as well as what they estimate, are

s^2_{w}
the symbol for within-groups variance estimate; $\frac{SS_{\text{w}}}{df_{\text{w}}}$

WITHIN-GROUPS VARIANCE ESTIMATE

$$s^2_{\text{w}} = \frac{SS_{\text{w}}}{df_{\text{w}}} \xrightarrow{\text{estimates}} \sigma^2 \text{ (inherent variance, sometimes called the error variance)} \qquad (18.6)$$

between-groups variance estimate an estimate of inherent variance plus treatment effect (if any exists)

BETWEEN-GROUPS VARIANCE ESTIMATE

$$s^2_{\text{bet}} = \frac{SS_{\text{bet}}}{df_{\text{bet}}} \xrightarrow{\text{estimates}} \sigma^2 + \textit{treatment effect} \qquad (18.7)$$

s^2_{bet} the symbol for between-groups (treatments) variance estimate; $\frac{SS_{\text{bet}}}{df_{\text{bet}}}$

Statisticians often refer to s^2_{w} as the *mean square within* (or sometimes the *mean square error*) and s^2_{bet} as the *mean square between*, symbolized MS_{w} and MS_{bet}, respectively. Although we will continue to use the former designations, you should be familiar with these alternative terms because they are often used in other books and articles.

By now you are probably asking what the variation in scores within and among groups has to do with testing hypotheses about means. The answer is that these two kinds of variation provide the basis for testing H_0: $\mu_A = \mu_B = \mu_C$. If H_0 is true (i.e., if there is no treatment effect), the sample means will tend to cluster closely around $\overline{\overline{X}}$, and s^2_{bet} will be an unbiased estimate of the inherent variation, σ^2. It will, therefore, estimate the same quantity as that estimated by s^2_{w}, and the two should be equal within the limits of sampling variation. On the other hand, if H_0 is false (if there is a treatment effect), the sum of squares of the deviations of the $\overline{X}$s around $\overline{\overline{X}}$ will tend to be larger, and s^2_{bet} will tend to be larger than s^2_{w}. The test of H_0 determines whether s^2_{bet} is so much larger than s^2_{w} that it exceeds the limits of sampling variation.

F ratio a ratio of within- and between-groups variance estimates. $\frac{s^2_{\text{bet}}}{s^2_{\text{w}}} = \frac{\text{inherent variance} + \text{treatment effect}}{\text{inherent variance}}$ When there is no treatment effect, the ratio should approximate 1.0

To complete the analysis of variance, we require a method of comparing s^2_{bet} with s^2_{w}. This is the ***F* ratio**, a statistic named in honor of Ronald Fisher. The *F* ratio is formed by the ratio of the two independent variance estimates:

F RATIO FOR ONE-WAY ANALYSIS OF VARIANCE

$$F = \frac{s^2_{\text{bet}}}{s^2_{\text{w}}} \qquad (18.8)$$

In constructing this ratio, s^2_{bet} is *always* placed in the numerator and s^2_{w} is always placed in the denominator. If H_0 is true and there is no treatment effect, s^2_{bet} and s^2_{w} will both be estimates of the inherent variance only, and the ratio should be *about* 1.0:

$$F = \frac{\textit{inherent variance} + \textit{treatment effect}}{\textit{inherent variance}}$$

If H_0 is false, s^2_{bet} will estimate the inherent variance plus the treatment effect and the ratio should exceed 1.0. The greater the treatment effect, the more *F* tends to exceed unity.

F distribution a theoretical relative frequency distribution of all the values of *F* that would be obtained by chance from an infinite number of samples of a particular size drawn from given populations

When H_0 is true and certain assumptions hold (see Section 18.11), *F* ratios follow the theoretical *F* distribution presented in Table E of Appendix D. Like the *t* distribution, the ***F* distribution** is actually a family of curves depending on the degrees of freedom. Figure 18.5 shows the distribution of *F* for 4 degrees of freedom in the numerator and 20 degrees of freedom in the denominator. Note that this distribution is positively skewed. This is intuitively reasonable. If the estimate of σ^2 in the numerator is smaller than that in the denominator, *F* will be less than 1.00 but not less than zero (because variance estimates are never negative). But if the estimate in the numerator is larger than that in the denominator, the *F* ratio may be much larger than 1.00. *The null hypothesis of equality of population means will be rejected only if the calculated value of* F *is larger than that expected through random sampling if the hypothesis is true.* Consequently, the region of rejection is entirely in the upper tail of the *F* distribution. (Recall, however, that the *F* test is nondirectional—see Section 18.1.)

True to its name, analysis of variance is concerned with the variance as the basic measure of variation. However, keep in mind that although we use variances to make decisions, ANOVA is nonetheless a test about a null hypothesis of differences among means.

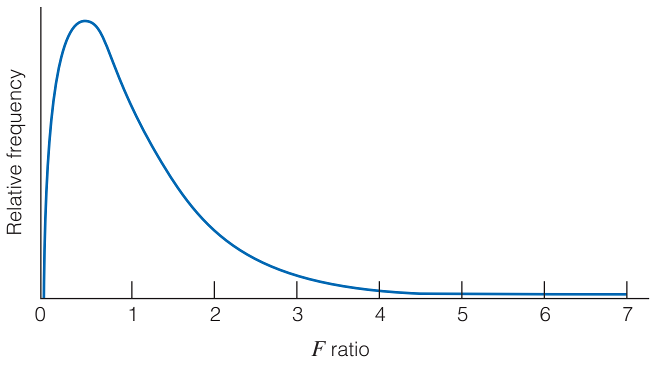


FIGURE 18.5 Distribution of F for 4 degrees of freedom in the numerator and 20 degrees of freedom in the denominator.

TABLE 18.1 *Summary Table of One-Way ANOVA, Showing the Partition of Sums of Squares, the Degrees of Freedom, the Variance Estimates, and* F

SOURCE	*SS*	*df*	s^2	*F*
Between groups	Formula 18.5	$k - 1$	$\frac{SS_{\text{bet}}}{df_{\text{bet}}}$	$\frac{s^2_{\text{bet}}}{s^2_{\text{w}}}$
Within groups	Formula 18.4	$n_{\text{total}} - k$	$\frac{SS_{\text{w}}}{df_{\text{w}}}$	
Total	Formula 18.3*a*	$n_{\text{total}} - 1$		

18.6 The Summary Table

In analysis of variance, it is convenient (and customary) to present a summary table indicating the source of variation, the corresponding sum of squares, the associated degrees of freedom, the variance estimate, and the calculated value of F. Table 18.1 gives the general form of a summary table for one-way analysis of variance. One advantage of such a table is that it shows the basic result of the analysis very clearly.

The first column lists the sources of variation. Note that the between-groups variation is listed above within-groups variation, thus setting up the F ratio. The sums of squares are listed in the second column, and their respective degrees of freedom in the third. As a check on your work, $SS_{\text{bet}} + SS_{\text{w}}$ should equal SS_{total}, and $df_{\text{bet}} + df_{\text{w}}$ should equal df_{total}. The numerator (s^2_{bet}) and denominator (s^2_{w}) of the F ratio, obtained by dividing SS by df, are shown in the fourth column. Notice that $s^2_{\text{bet}} + s^2_{\text{w}}$ does *not* equal s^2_{total}.

Table 18.1 does not show s^2_{total}. This quantity cannot be used in ANOVA because the test of H_0 requires the two variance estimates being compared to be independent of one another. Because SS_{total} contains SS_{w} and SS_{bet} ($SS_{\text{w}} + SS_{\text{bet}} = SS_{\text{total}}$), a variance estimate based on SS_{total} would obviously not be independent of either s^2_{w} or s^2_{bet}.

18.7 Example: Does Playing Violent Video Games Desensitize People to Real-Life Aggression?

Studies have shown that playing violent video games for as little as 20 minutes can result in physiological desensitization to aggression (e.g., decreased heart rate when watching actual violence) (Carnagey, Anderson, & Bushman, 2007). But does it change behavior? Bushman and Anderson (2009) recently showed that when college students played a violent video game they were slower to go to help the victim of a staged fight (between actors) compared to students who had just played a nonviolent video game.

Suppose we wanted to replicate this study at our college and include in the study a third group of students who were not playing any video games. We randomly assign 15 students from our college to three groups: (A) 5 who read a magazine for 20 minutes, (B) 5 who watch a nonviolent video game for 20 minutes, and (C) 5 who watch a violent video game for 20 minutes. Three minutes after the 20-minute session, we have two professional actors stage a fight outside the lab and one pretends to have a badly hurt ankle. We then record the amount of time it takes our subjects to go help the "injured" victim. (The results, in seconds, are hypothetical.) The null hypothesis is

$$H_0: \mu_A = \mu_B = \mu_C$$

If any one group differs from the other two or if all three are different, we will reject H_0 and accept H_A.

Due to an emergency phone call, one of the subjects in Group A leaves in the middle of the study, resulting in two groups with five subjects each and a third group with four subjects. The final results are as follow:

Time (in seconds)

GROUP *A*	**GROUP *B***	**GROUP *C***
17	12	73
10	15	83
28	25	64
13	30	52
	18	78

We first calculate each sample mean and the grand mean:

$$\overline{X}_A = \frac{68}{4} = 17 \qquad \overline{X}_B = \frac{100}{5} = 20 \qquad \overline{X}_C = \frac{350}{5} = 70$$

$$\overline{\overline{X}} = \frac{518}{14} = 37$$

Next, we calculate SS_{w} and SS_{bet}:

$$SS_{\text{w}} = \sum^{\substack{\text{all}\\\text{scores}}} (X - \overline{X})^2$$

$$= \sum(X_A - \overline{X}_A)^2 + \sum(X_B - \overline{X}_B)^2 + \sum(X_C - \overline{X}_C)^2$$

$$= \begin{bmatrix} (17-17)^2 \\ (10-17)^2 \\ (28-17)^2 \\ (13-17)^2 \end{bmatrix} + \begin{bmatrix} (12-20)^2 \\ (15-20)^2 \\ (25-20)^2 \\ (30-20)^2 \\ (18-20)^2 \end{bmatrix} + \begin{bmatrix} (73-70)^2 \\ (83-70)^2 \\ (64-70)^2 \\ (52-70)^2 \\ (78-70)^2 \end{bmatrix}$$

$$= 186 \qquad + \qquad 218 \qquad + \qquad 602$$

$$= 1006$$

$$
\begin{aligned}
SS_{\text{bet}} &= \sum^{k} n_i(\overline{X}_i - \overline{\overline{X}})^2 \\
&= 4(17 - 37)^2 + 5(20 - 37)^2 + 5(70 - 37)^2 \\
&= 1600 + 1445 + 5445 \\
&= 8490
\end{aligned}
$$

Although it is not required for the F ratio, as a check on our work we also calculate SS_{total}:

$$
\begin{aligned}
SS_{\text{total}} &= \sum (X - \overline{\overline{X}})^2 \\
&= (17 - 37)^2 + (10 - 37)^2 + (28 - 37)^2 + (13 - 37)^2 \\
&\quad + (12 - 37)^2 + (15 - 37)^2 + (25 - 37)^2 + (30 - 37)^2 + (18 - 37)^2 \\
&\quad + (73 - 37)^2 + (83 - 37)^2 + (64 - 37)^2 + (52 - 37)^2 + (78 - 37)^2 \\
&= 9496
\end{aligned}
$$

Because $SS_{\text{total}} = SS_{\text{w}} + SS_{\text{bet}}$, we can be confident that our calculations are correct. With each calculation, we enter the result in a summary table, shown in Table 18.2.

We then calculate the degrees of freedom:

$$
\begin{aligned}
df_{\text{bet}} &= k - 1 = 3 - 1 = 2 \\
df_{\text{w}} &= n_{\text{total}} - k = 14 - 3 = 11 \\
df_{\text{total}} &= n_{\text{total}} - 1 = 14 - 1 = 13
\end{aligned}
$$

As a check on our work, we see that $df_{\text{total}} = df_{\text{bet}} + df_{\text{w}}$. With the sums of squares and their respective degrees of freedom, we now calculate s^2_{bet} and s^2_{w}:

$$
s^2_{\text{bet}} = \frac{SS_{\text{bet}}}{df_{\text{bet}}} = \frac{8490}{2} = 4245
$$

$$
s^2_{\text{w}} = \frac{SS_{\text{w}}}{df_{\text{w}}} = \frac{1006}{11} = 91.45
$$

The F ratio is already set up in the summary table:

$$
F = \frac{s^2_{\text{bet}}}{s^2_{\text{w}}} = \frac{4245}{91.45} = 46.42
$$

In this example, $df_{\text{bet}} = 2$ and $df_{\text{w}} = 11$. To obtain the critical value of F, turn to Table E in Appendix D and locate the entries at the intersection of 2 *df* for the numerator and 11 *df* for the denominator. (Be careful not to look for the value corresponding to 11 *df* for the numerator and 2 *df* for the denominator; the critical value of F is not the same!) If we had adopted the .05 level of significance, the critical value of F is 3.98 (or 7.20 for the .01 level). If H_0 is true, we would obtain a sample F ratio greater than 3.98 only 5% of the time through

TABLE 18.2 *Summary Table of Analysis of Variance for the Test of the Hypothesis* H_0: $\mu_A = \mu_B = \mu_C$

SOURCE	*SS*	*df*	s^2	*F*
Between groups	8490	2	4245	46.42
Within groups	1006	11	91.45	
Total	9496	13		

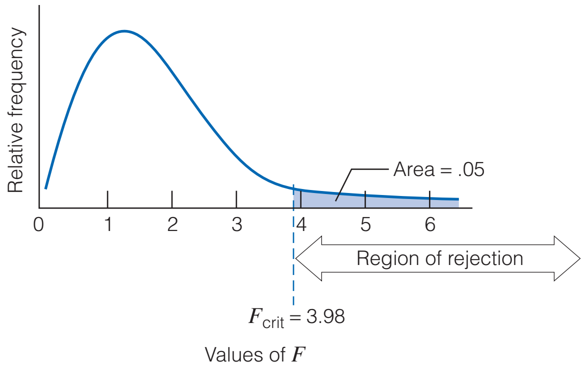


FIGURE 18.6 Sampling distribution of F for 2 and 11 degrees of freedom; $\alpha = .05$.

random sampling. It is clear that the obtained $F = 46.42$ falls beyond the critical value, so we reject H_0. The F distribution for 2 and 11 degrees of freedom and the region of rejection are shown in Figure 18.6.

Because our calculated value of F exceeds the critical value, we conclude that the population means differ and, therefore, that there is a real difference among the three conditions. It is tempting at this point to scan the sample data and say that students who play violent video games are slower to help victims of real-life violence. However, when the obtained F exceeds the critical value of F, *we can conclude only that there is a difference among groups.* To determine the manner in which the groups differ, we must conduct additional tests. We will discuss some of these tests in Section 18.14.

18.8 Comparison of *t* and *F*

ANOVA is closely related to the t test and, for the special case of two groups, leads to exactly the same conclusions as the t test. In fact, you can use ANOVA instead of t in the two-sample, independent-groups design. Recall from Chapter 14 that we summarized the t test for a difference between two independent means as

$$t = \frac{\begin{matrix}\textit{obtained difference} \\ \textit{between sample means}\end{matrix} - \begin{matrix}\textit{hypothesized difference between} \\ \textit{population means (when } \mathrm{H}_0 \textit{ is true)}\end{matrix}}{\textit{estimated standard error of the difference between two means}}$$

In this formula, the estimated standard error of the mean ($s_{\overline{X}-\overline{Y}}$) is the difference among scores due to chance (no treatment effect).

Now, look at the formula for F. For two samples,

$$F = \frac{\textit{variance estimate based on the difference between sample means}}{\textit{variance estimate within samples}}$$

The denominator is again the difference among scores due to chance (no treatment effect), except that it is an estimate of the variance whereas the denominator for the t formula is an estimate of a standard deviation. If we square a standard deviation, we get back to the variance. In fact, *with two samples*:

$F = t^2$ relationship between F and t for a two-sample, independent-groups design

RELATIONSHIP BETWEEN F AND t

$$F = t^2 \tag{18.9}$$

"FIND OUT WHO SET UP THIS EXPERIMENT. IT SEEMS THAT HALF OF THE PATIENTS WERE GIVEN A PLACEBO, AND THE OTHER HALF WERE GIVEN A DIFFERENT PLACEBO."

© Sidney Harris, ScienceCartoonsPlus.com

Try using ANOVA with the sample problem in Section 14.5, for which t was calculated to be 2.074. You will find that $F = 2.074^2 = 4.30$.

18.9 Raw-Score Formulas for Analysis of Variance

In Sections 18.3 through 18.6, we developed analysis of variance using the sums of squares, obtained by computing each deviation score and squaring it and then summing the results. This approach is best for understanding ANOVA. However, it is doubtful that you would ever use this method to analyze data in the real world. Today, most people use computers to do this work (see the next section). If you have to compute F by hand, and the goal is ease of computation rather than understanding, sums of squares are more easily computed from raw scores. We give the raw score formulas for SS_{w}, SS_{bet}, and SS_{total} in this section and illustrate their use with the data from Section 18.7. We start by obtaining the values of ΣX and ΣX^2 for each group separately. All other computations flow from these basic quantities and the sample sizes.

$$\text{Group } A\text{: } n_A = 4$$

$$\sum X_A = 17 + 10 + 28 + 13 = 68$$

$$\sum X_A^2 = 289 + 100 + 784 + 169 = 1342$$

$$\text{Group } B\text{: } n_B = 5$$

$$\sum X_B = 12 + 15 + 25 + 30 + 18 = 100$$

$$\sum X_B^2 = 144 + 225 + 625 + 900 + 324 = 2218$$

$$\text{Group } C\text{: } n_C = 5$$

$$\sum X_C = 73 + 83 + 64 + 52 + 78 = 350$$

$$\sum X_C^2 = 5329 + 6889 + 4096 + 2704 + 6084 = 25{,}102$$

Next, we sum the values of ΣX, ΣX^2, and n across the groups to obtain

$$\sum^{\substack{\text{all}\\\text{scores}}} X = \sum X_A + \sum X_B + \sum X_C = 68 + 100 + 350 = 518$$

$$\sum^{\substack{\text{all}\\\text{scores}}} X^2 = \sum X_A^2 + \sum X_B^2 + \sum X_C^2 = 1342 + 2218 + 25{,}102 = 28{,}662$$

$$n_{\text{total}} = n_A + n_B + n_C = 4 + 5 + 5 = 14$$

Now we have everything needed to obtain SS_{w}, SS_{bet}, and SS_{total}:

WITHIN-GROUPS SUM OF SQUARES:
RAW-SCORE FORMULA

$$SS_{\text{w}} = \sum^{\substack{\text{all}\\\text{scores}}} X^2 - \left[\frac{\left(\sum X_A\right)^2}{n_A} + \frac{\left(\sum X_B\right)^2}{n_B} + \cdots\right] \tag{18.10}$$

For our example,

$$\begin{aligned} SS_{\text{w}} &= 28{,}662 - \left[\frac{(68)^2}{4} + \frac{(100)^2}{5} + \frac{(350)^2}{5}\right] \\ &= 28{,}662 - [1156 + 2000 + 24{,}500] \\ &= 28{,}662 - 27{,}656 \\ &= 1006 \end{aligned}$$

BETWEEN-GROUPS SUM OF SQUARES:
RAW-SCORE FORMULA

$$SS_{\text{bet}} = \left[\frac{\left(\sum X_A\right)^2}{n_A} + \frac{\left(\sum X_B\right)^2}{n_B} + \cdots\right] - \frac{\left(\sum^{\substack{\text{all}\\\text{scores}}} X\right)^2}{n_{\text{total}}} \tag{18.11}$$

$$\begin{aligned} SS_{\text{bet}} &= \left[\frac{(68)^2}{4} + \frac{(100)^2}{5} + \frac{(350)^2}{5}\right] - \frac{(518)^2}{14} \\ &= 27{,}656 - 19{,}166 \\ &= 8490 \end{aligned}$$

TOTAL SUM OF SQUARES: RAW-SCORES FORMULA

$$SS_{\text{total}} = \sum^{\substack{\text{all}\\\text{scores}}} X^2 - \frac{\left(\sum^{\substack{\text{all}\\\text{scores}}} X\right)^2}{n_{\text{total}}} \tag{18.12}$$

$$SS_{\text{total}} = 28{,}662 - \frac{(518)^2}{14}$$
$$= 28{,}662 - 19{,}166$$
$$= 9496$$

Notice that the values for SS_w, SS_{bet}, and SS_{total} are exactly the same as those we obtained using the deviation-score formulas in Section 18.7. The completion of ANOVA from this point on is done just as in that section.

18.10 Calculation of ANOVA for Independent Measures with IBM SPSS

In this section, you will learn how to use IBM SPSS to calculate the *F* statistic to test hypotheses about more than two independent means. You will use the data from the example in Section 18.7.

In the **Data Editor**, all the scores for the dependent variable (i.e., time) should be entered in a single column. Now, in the next column, similar to Section 14.7, we need to identify from what condition the scores came. We will enter 1s for scores of subjects in Group A, 2s for scores of subjects in Group B, and 3s for scores of subjects in Group C. In the **Variable View**, for our dependent variable, we will use Time as the variable name and Time to Help as the label. For the second variable, which was used to indicate group membership, we will use Condition for the variable name and Experimental Condition as the label. Again, similar to Section 14.7, we will use the column labeled Values. Now, for the variable called Condition, click on Values. A dialog box called Value Labels will appear. In the Value box, type "1" and, in the Label box, type "Group A." Click Add. Now, in the Value box, type "2" and, in the Label box, type "Group B." Click Add. Now, in the Value box, type "3" and, in the Label box, type "Group C." Click OK.

To conduct a hypothesis test on more than two independent means in IBM SPSS, click **Analyze,** then select **General Linear Model** followed by **Univariate**. Figure 18.7 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze → General Linear Model → Univariate**.

Now, a dialog box will appear. Move the Time to Help variable (i.e., Time) to the Dependent Variable box and the Experimental Condition variable (i.e., Condition) to the Fixed Factor(s) box. Figure 18.8 shows the dialog box after the variables have been correctly moved. Click OK.

Table 18.3 shows the IBM SPSS output. After the syntax, there is a section of output labeled Between-Subjects Factors. As you can see, it shows the sample size in each of the three conditions. The next section of output is labeled Tests of Between-Subjects Effects. In the column labeled Source, we will focus on the rows labeled Condition, Error, and Corrected Total. As you will see under the column labeled Type III Sum of Squares, the sum of squares for Condition, Error, and Corrected Total, respectively, are $SS_{bet} = 8490.00$, $SS_W = 1006.00$, and $SS_{total} = 9496$. (Recall from Section 18.3 that SS_W is also referred to as SS_{error}.) The distinction regarding Type III Sum of Squares is irrelevant when dealing with a single independent variable. Hays (1994) provides a detailed description of the different types that are relevant for designs discussed in Chapter 19 when sample sizes within conditions are not equal. In the next column, the *df* are the same as those presented in Table 18.2. The column labeled Mean Square presents the same variance estimates as those found in Table 18.2. As we noted in Section 18.5, mean square is another term often used to represent the different variance estimates. Finally, we see that the *F* ratio is 46.417 and, under Sig., the *p*-value associated for our *F*

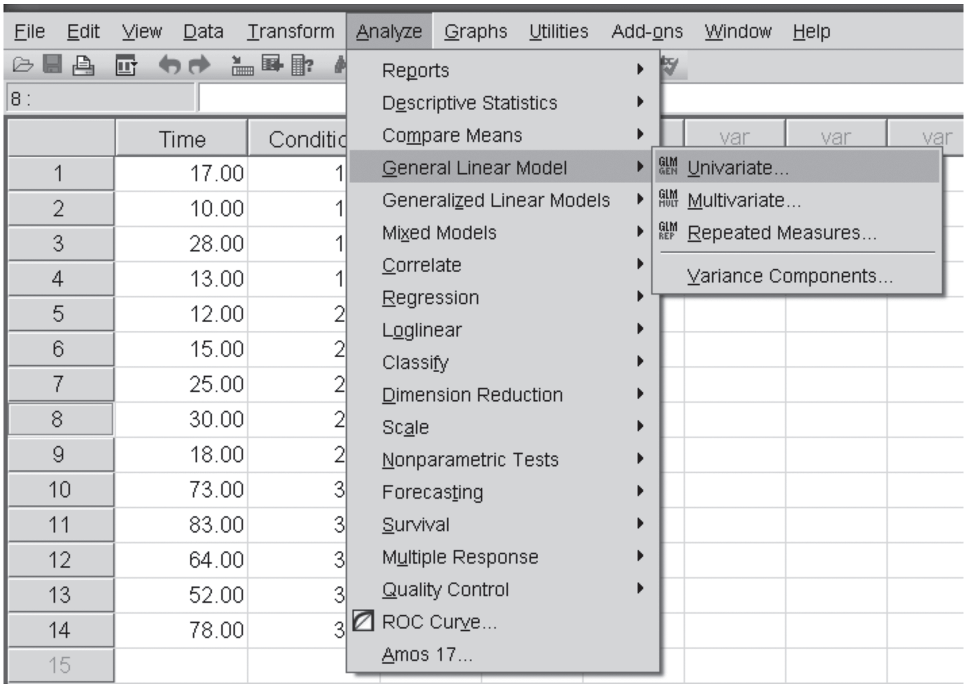
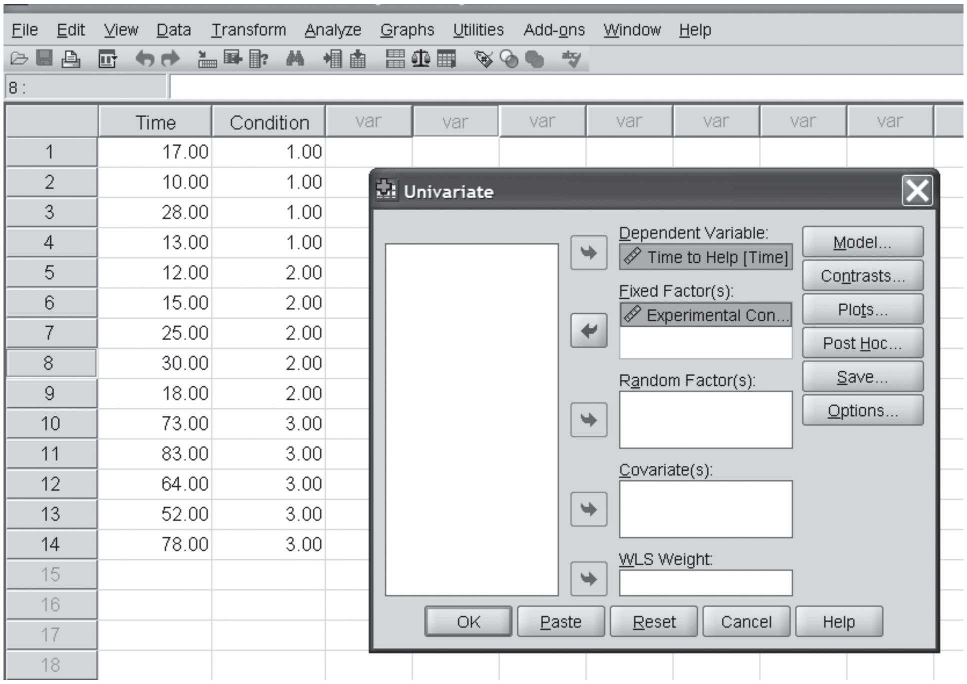

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
8 :
Time
Conditio
var var var
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Univariate...
Multivariate...
Repeated Measures...
Variance Components...

FIGURE 18.7 IBM SPSS menus when calculating ANOVA.



	Time	Condition
1	17.00	1.00
2	10.00	1.00
3	28.00	1.00
4	13.00	1.00
5	12.00	2.00
6	15.00	2.00
7	25.00	2.00
8	30.00	2.00
9	18.00	2.00
10	73.00	3.00
11	83.00	3.00
12	64.00	3.00
13	52.00	3.00
14	78.00	3.00

Univariate

Dependent Variable: Time to Help [Time]

Fixed Factor(s): Experimental Con...

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
8 :
var var var var var var var
15
16
17
18
Random Factor(s):
Covariate(s):
WLS Weight:
Model...
Contrasts...
Plots...
Post Hoc...
Save...
Options...
OK
Paste
Reset
Cancel
Help

FIGURE 18.8 IBM SPSS dialog box after selecting variables for a one-way ANOVA.

TABLE 18.3 *IBM SPSS Output from a One-Way ANOVA*

```
UNIANOVA Time By Condition
  /METHOD=SSTYPE (3)
  /INTERCEPT=INCLUDE
  /CRITERIA=ALPHA (0.05)
  /DESIGN=Condition.
```

Between-Subjects Factors

		Value Label	*N*
Experimental Condition	1.00	Group A	4
	2.00	Group B	5
	3.00	Group C	5

Tests of Between-Subjects Effects
Dependent Variable: Time to Help

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	8490.000[a]	2	4245.000	46.417	.000
Intercept	17613.846	1	17613.846	192.597	.000
Condition	8490.000	2	4245.000	46.417	.000
Error	1006.000	11	91.455		
Total	28662.000	14			
Corrected Total	9496.000	13			

[a] R Squared = .894 (Adjusted R Squared = .875)

ratio with $df_{\text{bet}} = 2$ and $df_{\text{w}} = 11$ is less than .001. Remember that you should not be interpreting the Sig. as $p = 0$ since IBM SPSS only displays the first three numbers to the right of the decimal. However, if you would like to see the actual *p*-value, you can double-click on the table and select the cell with the *p*-value.

18.11 Assumptions Associated with ANOVA

Because $F = t^2$, you should not be surprised to learn that the assumptions underlying the F test for one-way ANOVA are the same as those for the t test for independent samples (see Section 14.11):

1. The populations are normally distributed.
2. The variances of the several populations are the same (homogeneity of variance).
3. Selection of elements comprising any particular sample is independent of selection of elements of any other sample.
4. Samples are drawn at random with replacement.

As with the t test for independent means, *moderate* departure from the normal distribution specified in the first assumption does not unduly disturb the outcome of the test (see Glass, Peckham, & Sanders, 1972). This is especially true as sample size increases. However, with highly skewed populations, ANOVA results in less accurate probability statements (Tomarken & Serlin, 1986). (Again, see the Point of Controversy in Chapter 6.) When samples are quite small and there is serious question about the assumption of normality, one possible alternative is the Kruskal–Wallis nonparametric one-way ANOVA (see Section 21.8).

Violations of the second assumption, homogeneity of variance, are also negligible *when sample sizes are equal*. Heterogeneity of variance ordinarily becomes a problem only when $\sigma^2_{\text{largest}} = 4\sigma^2_{\text{smallest}}$ or, with more moderate differences, when sample sizes differ considerably (Milligan, Wong, & Thompson, 1987; Tomarken & Serlin, 1986). If there is a choice, you should try to select samples of equal size for each treatment condition. This minimizes the effect of failing to satisfy the condition of homogeneity of variance.

The third assumption, independence of samples, reminds us that the ANOVA procedure described to this point is not appropriate for repeated measures on the same subjects nor when matched subjects have been assigned to treatment groups. We will introduce ANOVA for repeated measures in Section 18.18.

The biggest problem is probably that of obtaining random samples. In our hypothetical problem in Section 18.7, we used random sampling, but in the real world most researchers use random assignment of available subjects. Random sampling is more of a premise than a requirement—it allows us to extend conclusions to the population. Recall our earlier discussions of this problem in Sections 14.12 and 14.13.

18.12 Effect Size

In Section 14.9, you learned about measures of effect size that may be used for tests of a hypothesis about the difference between two independent sample means. Recall that there are two families of effect size: the *d* family (expressing the difference between the two means relative to the standard deviation of the variable) and the *r* family (expressing effect size as a correlation coefficient).

In this section, you will learn about two measures of effect size commonly used with ANOVA for independent groups. Both are members of the *r* family. The first is **eta-squared (η^2)**, which is also known as the correlation ratio:

eta-squared (η^2) an *r* family measure of effect size for independent-groups ANOVA $\dfrac{SS_{\text{bet}}}{SS_{\text{total}}}$

FORMULA FOR ETA-SQUARED

$$\eta^2 = \frac{SS_{\text{bet}}}{SS_{\text{total}}} \tag{18.13}$$

In brief, η^2 provides a measure of the strength of association between the independent and dependent variables. More specifically, it gives us the proportion of the total variance due to the independent variable (Section 9.5). J. Cohen (1988) has suggested that for ANOVA, .01 = small effect, .06 = medium effect, and .14 = large effect (but recall from Section 13.2 that importance can be judged relative only to the circumstances of the particular study). In our example in Section 18.7, $\eta^2 = 8490/9496 = .894$, an unusually high effect size (remember, our example was hypothetical). Notice at the bottom of Table 18.3 that IBM SPSS reports η^2 as R squared = .894.

Eta-squared is simple, but it is also biased in an upward direction. A better measure of effect size for ANOVA is **omega-squared ($\hat{\omega}^2_{\text{bet}}$)** (Fleiss, 1969; Vaughan & Corballis, 1969). Omega-squared provides a relatively unbiased estimate of the strength of association—it estimates the proportion of the variance in the dependent variable (in the population) that is due to the *k* levels of treatment. The formula is:

omega-squared ($\hat{\omega}^2_{\text{bet}}$) an *r* family measure of effect size for independent-groups ANOVA $\dfrac{SS_{\text{bet}} - (k-1)(s^2_{\text{w}})}{SS_{\text{total}} + s^2_{\text{w}}}$

$$\hat{\omega}^2_{\text{bet}} = \frac{\sigma^2_{\text{bet}}}{\sigma^2_{\text{bet}} + \sigma^2_{\text{within}}}$$

Of course, we do not have the actual population variances, so we have to estimate omega-squared:

FORMULA FOR OMEGA-SQUARED

$$\hat{\omega}^2_{\text{bet}} = \frac{SS_{\text{bet}} - (k - 1)(s^2_{\text{w}})}{SS_{\text{total}} + s^2_{\text{w}}} \tag{18.14}$$

or, calculating $\hat{\omega}^2_{\text{bet}}$ directly from the F ratio:

$$\hat{\omega}^2_{\text{bet}} = \frac{(k - 1)(F - 1)}{(k - 1)(F - 1) + (k)(n_{\text{cell}})}$$

As with η^2, J. Cohen (1988) suggests that we interpret $\hat{\omega}^2$ as follows: .01 = small effect, .06 = medium effect, and .14 = large effect. As an example, we will again use the data in Section 18.7:

$$\hat{\omega}^2_{\text{bet}} = \frac{8490 - (2)(91.45)}{9496 + 91.45}$$

$$= .866$$

Notice that this outcome is slightly lower than that obtained with η^2. This is due to η^2 being an overestimate of effect size.

In ANOVA with more than two groups, there are actually several correlation indices that may be used to measure effect size. For a discussion of each, see Rosnow, Rosenthal, and Rubin (2000).

18.13 ANOVA and Power

The factors affecting the power (the probability of rejecting a false null hypothesis) of the one-way analysis of variance are very similar to those affecting the *t* test (see Sections 13.4–13.10 and 14.10). The major influences are the following:

1. The actual difference among the population means.
2. The variance that is not attributable to treatment effect.
3. The degrees of freedom in the numerator.
4. The degrees of freedom in the denominator.
5. The level of significance (the probability of a Type I error).

As with hypotheses about two means, we can estimate power and sample sizes for tests of hypotheses about three or more means. If you wish to learn how to do this, a good reference is J. Cohen (1988).

18.14 *Post Hoc* Comparisons

Suppose that in our example concerning video games (Section 18.7) the obtained F ratio turned out to be *smaller* than the critical value of 3.98. We would conclude that our data do not support the presence of treatment effects, and that would be that. There would be no justification for further statistical work.

Such, however, was not the case. Our obtained F of 46.42 led us to reject H_0: $\mu_A = \mu_B = \mu_C$ in favor of the very broad alternative hypothesis that the three population means differed in *some* way. But where is the real difference (or differences)? All three might differ from one another, or two of them might be equal and differ from the third. To answer this question we may, on obtaining a significant overall F, proceed with further statistical comparisons involving the group means.

***post hoc* comparisons** statistical tests performed after obtaining a significant value of F that indicate which means are significantly different

The tests we use to make these specific comparisons are called ***post hoc* comparisons** or *a posteriori* (Latin, meaning "what comes later") comparisons. They keep us from taking undue advantage of chance. If we are comparing many means, we would expect some of the differences to be fairly substantial as a result of random sampling variation alone, even if the null hypothesis is true. Applying the conventional t test just to the largest or to all of the possible differences between sample means would substantially increase the probability of making a Type I error—that is, rejecting H_0 when it is true. Instead of using a sampling distribution that compares the means of only two samples (as is done with t), *post hoc* tests generally employ sampling distributions that compare the means of many samples. In short, *post hoc* tests protect us from making too many Type I errors by requiring a bigger difference (between sample means) before we can declare that difference to be statistically significant.

There are several commonly used *post hoc* tests. The only real difference among them is that some are more conservative than others. In ascending level of conservativeness with regard to Type I errors (but descending power), we could choose from among such tests as Duncan's multiple-range test, the Newman–Keuls test, Tukey's HSD test, and the Scheffé test. *To use any one of these tests, our* F *ratio must first be significant.*

Tukey's HSD test honestly significant difference; a commonly used *post hoc* test; critical HSD value = $q\sqrt{\frac{s_w^2}{n}}$

We will illustrate only one *post hoc* test, **Tukey's HSD test**. It does not inflate the probability of Type I error as much as many other tests, yet it is not nearly as conservative as the Scheffé test. HSD stands for "honestly significant difference." The test involves determining a critical HSD value for the data. Although the test can be used to make specific comparisons, most investigators make all possible pairwise comparisons. The hypothesis of equal population means is rejected for any pair of samples for which the difference between $\overline{X}$s is as large as or larger than the critical HSD value.

We determine the critical HSD value using the following formula:

Studentized range statistic (q) used with Tukey's HSD test; based on sampling distributions comparing many means

CRITICAL HSD VALUE FOR TUKEY'S HSD TEST

$$\text{HSD} = q\sqrt{\frac{s_w^2}{n}} \tag{18.15}$$

where: s_w^2 = the within-groups variance estimate
n = the number of cases in each group
q = the value of a new statistic called the **Studentized range statistic** (based on sampling distributions comparing many means rather than just two).

The value of q depends on the significance level (.05 or .01), the number of samples to be compared, and df_w. To find q, look in Table F in Appendix D. If the number of cases is not the same for all samples, you may use the harmonic mean, $\tilde{n}$, in place of n (Winer, Brown, & Michels, 1991). The harmonic mean is calculated as follows:

$\tilde{n}$
symbol for the
harmonic mean

HARMONIC MEAN FOR TUKEY'S HSD TEST

$$\tilde{n} = \frac{k}{(1/n_A) + (1/n_B) + \cdots + (1/n_k)} \tag{18.16}$$

where: k = the number of groups
n_k = the number of cases in the kth group.

To illustrate the use of Tukey's HSD test, let us return to our example from Section 18.7, in which we obtained a significant difference among the three video game conditions. We now wish to examine the pattern of differences among group means.

In our example, $df_{\mathrm{w}} = 11$, $k = 3$, and $\alpha = .05$. Entering Table F with these three values, we find that $q = 3.82$. Because the sample sizes were unequal, we must calculate the harmonic mean:

$$\tilde{n} = \frac{3}{(1/4) + (1/5) + (1/5)} = \frac{3}{.65} = 4.62$$

We already have $s_{\mathrm{w}}^2 = 91.45$. The critical HSD value is thus

$$\text{HSD} = 3.82\sqrt{\frac{91.45}{4.62}} = 17.0$$

Now we can compute the differences for all the possible pairs of sample means and display them. In the following table, each entry is the difference between the mean listed on the left and that listed at the top.

	$\overline{X}_A = 17$	$\overline{X}_B = 20$	$\overline{X}_C = 70$
$\overline{X}_A = 17$	0	−3	−53
$\overline{X}_B = 20$		0	−50
$\overline{X}_C = 70$			0

In this example, only the differences between $\overline{X}_c$ and the others exceed 17.0 and thus are significant at the .05 level. We can conclude that students who watch violent video games are slower to help a real-life victim of violence than students who watch nonviolent video games or no video games.

For an excellent review of the logic of *post hoc* (multiple) comparisons, as well as an update of the strengths and weaknesses of several different tests, see Curran-Everett (2000).

18.15 Some Concerns about *Post Hoc* Comparisons

Although most researchers still follow an ANOVA with *post hoc* tests (either tests of all possible pairs or all possible combinations), many leading statisticians caution about their use (see Wilkinson et al., 1999). Why? First, most researchers do not need to see the results of comparing all possible pairs to answer the question(s) they are interested in, and tests for *all* possible pairs are very conservative, thus sacrificing power. Second, the use of tests of all possible pairs often resembles a fishing expedition:

> Reviewers should require writers to articulate their expectations well enough to reduce the likelihood of *post hoc* rationalizations. Fishing expeditions are often recognizable by the promiscuity of their explanations. (Wilkinson and the Task Force on Statistical Inference, 1999)

In place of *post hoc* tests, these statisticians urge that researchers examine trends or clusters of effects (see Rosnow, Rosenthal, & Rubin, 2000, for these more advanced techniques). Most important, they advise, "If a specific contrast interests you, examine it" (Wilkinson et al., 1999). We will now look at how we might do this.

18.16 An Alternative to the *F* Test: Planned Comparisons

In our example about video games, suppose that we had not had a well-developed rationale for focusing on a particular comparison prior to beginning the study. Our study, like many of those in the real world, would have been a "fishing expedition": We would have wanted to know whether any of the video game conditions makes a difference in helping others and, if so, which one(s)? Only after finding (through the overall F test) that there was a difference did we reexamine the various possible comparisons. However, the price we pay for this approach, you recall, is loss of power; the differences among sample means must be larger (to be significant with *post hoc* tests) than with t tests to protect us from making Type I errors.

For example, consider a hypothetical experiment in which four new drugs are to be compared with a placebo. Suppose only one of the drugs is truly effective. If we compare it alone with the placebo, the end result might be $p < .01$. If we include two other groups tested with ineffective drugs, the overall end result might drop to $p < .05$. With all the ineffective drugs included, we might fail to reach a significant result at all (Sinclair, 1988). Thus, the failure to reject the null hypothesis is often due to the insensitivity of the F test.

There is an alternative if we know *in advance* what comparisons will interest us. If theory or the logic of the study (rather than the obtained results) suggests the comparisons to be made, we can use what are known as **planned comparisons** or *a priori comparisons*. Suppose, for example, that we want to compare the effectiveness of these three methods of instruction: lecture, lecture plus movie, and lecture plus audio tape. It might make sense to construct two comparisons: (1) lecture versus the average of the two methods that use supplementary material and (2) lecture plus movie versus lecture plus audio tape. The advantage of planning comparisons in advance is that a *significant overall* F *ratio is not required* and, in fact, we can avoid the omnibus F test altogether if these are the only comparisons we are interested in.

planned comparisons tests used in place of ANOVA when the researcher knows in advance which comparisons are of interest; also called *a priori comparisons*

To make planned comparisons, we use the t test for independent groups. But instead of calculating the standard error by pooling the variance of just two groups, we use s^2_{w}, which pools the data from all the groups and is thus a better estimate:

STANDARD ERROR OF A COMPARISON

$$s_k = \sqrt{s^2_{\mathrm{w}}\left(\frac{a_A^2}{n_A} + \frac{a_B^2}{n_B} + \cdots + \frac{a_k^2}{n_k}\right)} \qquad (18.17)$$

where: s^2_{w} = the variance estimate that constitutes the denominator of the overall F test
a_A = the coefficient of the mean of Group A, and so forth (we will explain this shortly)
n_A = the number of cases in Group A, and so forth.

s_k symbol for standard error of a comparison

The number of degrees of freedom associated with this standard error is the number associated with s_w^2: $\sum^{k}(n-1)$.

At this point you are probably saying to yourself, "Now wait a minute—you told us that the major reason for using ANOVA is to avoid multiple *t* tests." That remains true. If there were no constraints and we did all the possible comparisons, we would substantially increase the likelihood of a Type I error. However, there are *three constraints when using planned comparisons*:

orthogonal comparisons comparisons that are unique and do not overlap

1. All questions are decided in advance so that they are not influenced by the data (as with the decision of whether to use a one-tailed test).
2. All comparisons are **orthogonal comparisons**, which means that each is unique and none overlaps another.
3. All questions can be answered by $k-1$ or fewer comparisons.

18.17 How to Construct Planned Comparisons

Suppose a study consists of a control condition (*A*) and three experimental conditions (*B*, *C*, and *D*), with an equal number of subjects in each group. If we wish to compare the control condition with the average of the experimental conditions, we can express the desired comparison as:

$$\overline{X}_A \text{ vs. } \frac{\overline{X}_B+\overline{X}_C+\overline{X}_D}{3}$$

or

$$1\overline{X}_A \text{ vs. } \left(\tfrac{1}{3}\,\overline{X}_B+\tfrac{1}{3}\,\overline{X}_C+\tfrac{1}{3}\,\overline{X}_D\right)$$

The hypothesis we propose to test is, of course, about the difference at the population level:

$$H_0\text{: } 1\mu_A-\tfrac{1}{3}\mu_B-\tfrac{1}{3}\mu_C-\tfrac{1}{3}\mu_D=0$$

Notice the following:

1. In a comparison, two quantities are contrasted with each other. In our example, the mean of *A* is contrasted with the average of the means of the remaining conditions, *B*, *C*, and *D*.
2. Each term in the null hypothesis of the comparison is multiplied by a coefficient. In our example, $+1$ is the coefficient for μ_A and $-\frac{1}{3}$ is the coefficient for μ_B, μ_C, and μ_D.
3. The total of the positive coefficients equals the total of the negative coefficients.

Other contrasts are, of course, possible. If we wish to compare the average of *A* and *B* with the average of *C* and *D*, the contrast would be

$$\frac{\overline{X}_A+\overline{X}_B}{2} \text{ vs. } \frac{\overline{X}_C+\overline{X}_D}{2}$$

or more formally

$$\tfrac{1}{2}\overline{X}_A+\tfrac{1}{2}\overline{X}_B-\tfrac{1}{2}\overline{X}_C-\tfrac{1}{2}\overline{X}_D$$

Similarly, for $\overline{X}_A$ vs. $\overline{X}_B$, we have

$$1\overline{X}_A-1\overline{X}_B$$

and for a comparison of the average of *A* and *B* with *D*, we have

$$\tfrac{1}{2}\overline{X}_A+\tfrac{1}{2}\overline{X}_B-1\overline{X}_D$$

In general, a comparison, K, may be expressed as

A COMPARISON

$$K = a_A\overline{X}_A + a_B\overline{X}_B + \cdots + a_k\overline{X}_k \tag{18.18}$$

where: a_A, a_B, etc. = the coefficients for the several levels of treatment
k = the number of levels of the particular treatment.

If some levels are not included in the comparison, the coefficients of the means of these groups are assigned the value zero.

To sum up, a comparison is constructed as follows:

Step 1: Decide what conditions (levels of treatment) are to be compared.

Step 2: Select coefficients for the means of the treatment levels to be used in the comparison, assigning a coefficient of zero to means not involved in the comparison, and making certain that the algebraic sum of all the coefficients is zero.

With k groups, there are $k - 1$ possible orthogonal comparisons. Remember, each comparison must be unique and none can overlap another. *When there are an equal number of subjects in each group, two comparisons are orthogonal if the sum of the products of the corresponding coefficients is zero.* Look at Table 18.4, which shows some comparisons possible with five groups. Comparisons 1 and 2 are orthogonal $[(+1)(0) + (0)(+1) + (0)(0) + (0)(-1) + (-1)(0) = 0]$, as are comparisons 2 and 3 $[(0)(+1) + (+1)(0) + (0)(-1) + (-1)(0) + (0)(0) = 0]$, but comparisons 1 and 3 are not $[(+1)(+1) + (0)(0) + (0)(-1) + (0)(0) + (-1)(0) = 1]$. To meet the requirement of orthogonality, *each comparison must be orthogonal to every other comparison.* If we eliminate comparison 3 in our example, each of the remaining $k - 1$ comparisons is orthogonal with the remaining three.

When group sample sizes are equal, we can construct a set of $k - 1$ orthogonal comparisons in the following manner:

Step 1: Construct the first comparison using all the levels of the treatment. For example, if there are four groups, the first comparison might be

$$1\overline{X}_A \text{ vs. } (\tfrac{1}{3}\overline{X}_B + \tfrac{1}{3}\overline{X}_C + \tfrac{1}{3}\overline{X}_D)$$

Step 2: The second comparison must be constructed wholly from groups that fall on *one* side of the first comparison. In our example, only a comparison involving B, C, and D will

TABLE 18.4 *Five Comparisons that Are Possible with Five Groups. Only Comparisons 1, 2, 4, and 5 Are Orthogonal*

	GROUP MEANS				
	$\overline{X}_A$	$\overline{X}_B$	$\overline{X}_C$	$\overline{X}_D$	$\overline{X}_E$
Comparison	***Coefficients***				
1: $\mu_A - \mu_E$	$+1$	0	0	0	-1
2: $\mu_B - \mu_D$	0	$+1$	0	-1	0
3: $\mu_A - \mu_C$	$+1$	0	-1	0	0
4: $\frac{1}{2}\mu_B + \frac{1}{2}\mu_D - \frac{1}{2}\mu_A - \frac{1}{2}\mu_E$	$-\frac{1}{2}$	$+\frac{1}{2}$	0	$+\frac{1}{2}$	$-\frac{1}{2}$
5: $\frac{1}{4}\mu_A + \frac{1}{4}\mu_B + \frac{1}{4}\mu_D + \frac{1}{4}\mu_E - \mu_C$	$+\frac{1}{4}$	$+\frac{1}{4}$	-1	$+\frac{1}{4}$	$+\frac{1}{4}$

meet this criterion. Again, we must use *all* the available groups. We might, for example, choose

$$1\overline{X}_C \text{ vs. } (\tfrac{1}{2}\overline{X}_B + \tfrac{1}{2}\overline{X}_D)$$

Step 3: Construct the third comparison by applying the procedure of Step 2 to the comparison just obtained. In our example, only one comparison is now possible:

$$1\overline{X}_B \text{ vs. } 1\overline{X}_D$$

Notice that we now have a set of $k - 1 = 3$ comparisons constructed from the $k = 4$ treatment means.

You can construct alternative sets of comparisons in this way. For our example, one possible alternative set is

$$(\tfrac{1}{2}\overline{X}_A + \tfrac{1}{2}\overline{X}_B) \text{ vs. } (\tfrac{1}{2}\overline{X}_C + \tfrac{1}{2}\overline{X}_D)$$

$$1\overline{X}_A \text{ vs. } 1\overline{X}_B$$

$$1\overline{X}_C \text{ vs. } 1\overline{X}_D$$

Although comparisons within a particular set approach mutual independence, this cannot be said for comparisons selected from more than one set. Consequently, in consideration of the reasons for doing the study, we decide in advance which set would be most appropriate and choose comparisons from that set only.

Let us return to our example of the three methods of instruction: lecture alone, lecture plus movie, and lecture plus audio tape. We suggested that it might make sense to construct two comparisons: (1) lecture versus the average of the two methods that use supplementary material and (2) lecture plus movie versus lecture plus audio tape. Are these two comparisons orthogonal? To answer this, we must determine the appropriate coefficients:

	1 LECTURE	2 LECTURE + MOVIE	3 LECTURE + AUDIO TAPE
$\mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3$	+1	$-\frac{1}{2}$	$-\frac{1}{2}$
$\mu_2 - \mu_3$	0	+1	−1

We now determine the sum of the products of the corresponding coefficients: $(+1)(0) + (-\frac{1}{2})(1) + (-\frac{1}{2})(-1) = 0$. The two comparisons are indeed orthogonal.

If we compute t for all $k - 1$ orthogonal comparisons, the mean of the squared t values will equal the omnibus F ratio. *The use of planned comparisons is the most powerful method of discovering a difference between population means, if one exists.* Not only does the requirement of orthogonality minimize the chance of making a Type 1 error, but it *allows us to calculate the exact probability of a Type I error.* Even Fisher (1949) believed that a *priori* comparisons are justified:

> When the (F) test does not demonstrate significant differentiation, much caution should be used before claiming significance for special comparisons. Comparisons, which the experiment was designed to make, may, of course, be made without hesitation. (R.A. Fisher, 1949, p. 57)

18.18 Analysis of Variance for Repeated Measures

To this point in the chapter, we have considered only the comparison of independently drawn random samples. Suppose, however, that we wish to test different dosages of a new experimental drug for hyperactive children, and that to control for individual differences among subjects, we prefer to test three dosages in the same subjects. We introduced the *t* test for testing hypotheses about two dependent means (and also discussed the advantages of the dependent-groups design) in Chapter 15. However, just as is the case with independent groups, we should not use multiple *t* tests when comparing three or more dependent means because of the increased probability of a Type I error. Instead, we use **analysis of variance for repeated measures**.

ANOVA for repeated measures an analysis of variance procedure used when one sample is tested under two or more treatment conditions

The logic of one-way ANOVA for repeated measures is the same as that for independent measures: We partition the total variability in the data into two parts, within groups and between groups. In the ANOVA for repeated measures, SS_w is also partitioned into two parts: (1) $\boldsymbol{SS_{subjects}}$, a measure of the variability due to individual differences, and (2) $\boldsymbol{SS_{residual}}$, a measure of the variability due to random error or chance. Figure 18.9 shows the partitioning for ANOVA for repeated measures.

$\boldsymbol{SS_{subjects}}$ a measure of the variability due to individual differences

$\boldsymbol{SS_{residual}}$ a measure of the variability due to chance (excluding the individual differences)

We can measure the variability due to individual differences because, unlike in the ANOVA for independent measures, each subject is tested under all conditions. This allows us to record a subject's overall score (by adding together all his or her scores), which we then compare with other subjects' overall scores. If two subjects' overall scores differ, we attribute it to individual differences. In the case where two subjects' overall scores are the same but their scores differ across conditions, we attribute it to random error.

The formula for SS_{total} is the same as for one-way ANOVA with independent groups (Formula 18.3*a*):

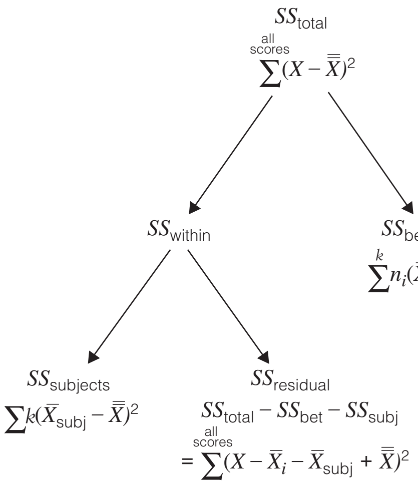
SS_{total}: DEVIATION-SCORES METHOD

$$SS_{total} = \sum^{\text{all scores}} (\overline{X} - \overline{\overline{X}})^2$$

The formula for SS_{bet} is also the same. Because each group has the same number of subjects in the repeated-measures design, we can modify the raw-scores formula for SS_{bet} (Formula 18.11).

SS_{bet}: DEVIATION-SCORES METHOD

$$SS_{bet} = \sum^{k} n_i(\overline{X}_i - \overline{\overline{X}})^2 \qquad (18.19a)$$


SS_{total}
$\sum^{\text{all scores}}(X - \overline{\overline{X}})^2$
SS_{within}
$SS_{between}$
$\sum^{k} n_i(\overline{X}_i - \overline{\overline{X}})^2$
$SS_{subjects}$
$\sum k(\overline{X}_{subj} - \overline{\overline{X}})^2$
$SS_{residual}$
$SS_{total} - SS_{bet} - SS_{subj}$
$= \sum^{\text{all scores}}(X - \overline{X}_i - \overline{X}_{subj} + \overline{\overline{X}})^2$

partition of the sums of squares with ANOVA for repeated measures
$SS_{total} = SS_{within} + SS_{between}$
$SS_{within} = SS_{subjects} + SS_{residual}$

FIGURE 18.9 Partitioning of the sums of squares for ANOVA for repeated measures.

SS_{bet}: RAW-SCORES METHOD FOR REPEATED MEASURES

$$SS_{\text{bet}} = \frac{\left(\sum X_A\right)^2 + \left(\sum X_B\right)^2 + \cdots}{n_k} - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X\right)^2}{n_{\text{total}}} \qquad (18.19b)$$

To calculate SS_{subjects}, use either of the following formulas:

SS_{subjects}: DEVIATION-SCORES METHOD

$$SS_{\text{subj}} = \sum k(\overline{X}_{\text{subj}} - \overline{\overline{X}})^2 \qquad (18.20a)$$

where: $\overline{X}_{\text{subj}}$ = a subject's mean score across all conditions
k = the number of conditions.

SS_{subjects}: RAW-SCORES METHOD

$$SS_{\text{subj}} = \frac{\sum X^2_{\text{subj}}}{k} - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X\right)^2}{n_{\text{total}}} \qquad (18.20b)$$

where: k = the number of conditions.

We find SS_{residual} by subtraction:

$$SS_{\text{resid}} = SS_{\text{total}} - SS_{\text{bet}} - SS_{\text{subj}} \qquad (18.21)$$

The degrees of freedom for SS_{subj} is the number of subjects minus one. For SS_{resid}, the degrees of freedom is $(df_{\text{subj}})(df_{\text{bet}})$.

To determine whether there is a treatment effect, we calculate the F ratio by dividing s^2_{bet} by s^2_{resid}:

$$F = \frac{s^2_{\text{bet}}}{s^2_{\text{resid}}} = \frac{\textit{random error} + \textit{treatment effect}}{\textit{random error}} \qquad (18.22)$$

***F* ratio for repeated-measures ANOVA**

$$\frac{s^2_{\text{bet}}}{s^2_{\text{resid}}} = \frac{\text{inherent variance} + \text{treatment effect}}{\text{inherent variance}}$$

When there is no treatment effect, the ratio should approximate 1.0.

Table 18.5 gives the general form of the summary table for ANOVA for repeated measures. Notice that only the F ratio for the independent variable is calculated. Rarely is anyone interested in knowing whether there is a significant difference among individuals.

The one-way ANOVA for repeated measures is a more powerful design than the one-way ANOVA for independent groups because it removes the variability due to individual differences (s^2_{bet} is divided by s^2_{resid}, which is smaller than s^2_{w}). The dependent-samples design also has some shortcomings, however (see Section 15.9).

To demonstrate ANOVA for repeated measures, let us return to our hypothetical example. We wish to test three levels of the independent variable: zero (Z), moderate (M), and high (H) dosages of a new experimental drug for hyperactive children. As the dependent variable,

TABLE 18.5 *Summary Table of One-Way ANOVA for Repeated Measures*

SOURCE	*SS*	*df*	s^2	*F*
Subjects	Formula 18.20*a* or 18.20*b*	$n - 1$	$\frac{SS_{\text{subj}}}{df_{\text{subj}}}$	—
Between groups	Formula 18.19*a* or 18.19*b*	$k - 1$	$\frac{SS_{\text{bet}}}{df_{\text{bet}}}$	$\frac{s^2_{\text{bet}}}{s^2_{\text{resid}}}$
Residual	Formula 18.21	$(df_{\text{subj}})(df_{\text{bet}})$	$\frac{SS_{\text{resid}}}{df_{\text{resid}}}$	
Total	Formula 18.3*a*	$(k)(n) - 1$		

we choose scores on a reading test that requires sustained attention for 30 minutes. The null and alternative hypotheses are

$$H_0\!: \mu_{\text{Z}} = \mu_{\text{M}} = \mu_{\text{H}}$$

$$H_{\text{A}}\!: \text{not } H_0$$

We randomly select 12 third graders from our school district and test them. Four of the children receive the zero dose first, followed by the moderate, and then the high dosage, with 1 week between tests. Four other children receive the drug in the order M, H, Z, and the remaining four are administered the drug in the order H, Z, M. Table 18.6 gives hypothetical results for our experiment, as well as the initial calculations. The sums of squarses are calculated as follows:

$$\begin{aligned} SS_{\text{total}} &= \sum^{\substack{\text{all}\\ \text{scores}}} (X - \overline{\overline{X}})^2 \\ &= (35 - 43.33)^2 + (23 - 43.33)^2 + \cdots + (33 - 43.33)^2 \\ &= 11{,}550 \end{aligned}$$

$$\begin{aligned} SS_{\text{bet}} &= \sum^{k} n_i(\overline{X}_i - \overline{\overline{X}})^2 \\ &= 12(33 - 43.33)^2 + 12(63 - 43.33)^2 + 12(34 - 43.33)^2 \\ &= 6{,}968.0 \end{aligned}$$

$$\begin{aligned} SS_{\text{subj}} &= \sum k(\overline{X}_{\text{subj}} - \overline{\overline{X}})^2 \\ &= 3(41.67 - 43.33)^2 + 3(32.67 - 43.33)^2 + \cdots + 3(41.00 - 43.33)^2 \\ &= 3{,}056.03 \end{aligned}$$

$$\begin{aligned} SS_{\text{resid}} &= SS_{\text{total}} - SS_{\text{bet}} - SS_{\text{subj}} \\ &= 11{,}550 - 6{,}968 - 3{,}056.03 \\ &= 1{,}525.97 \end{aligned}$$

Table 18.6 also gives the summary table for this example. Notice that $F = 50.23$. From Table E in Appendix D, we find that the critical value of F with 2 and 22 degrees of freedom

TABLE 18.6 *Data, Initial Calculations, and ANOVA Summary Table for the Test of the Hypothesis* H_0: $\mu_Z = \mu_M = \mu_H$, *Using Repeated Measures*

	READING SCORE FOR DOSAGE OF DRUG					
SUBJECT	*Zero*	*Moderate*	*High*	TOTAL (X_{subj})	X^2_{subj}	$\overline{X}_{\text{subj}}$
1	35	60	30	125	15,625	41.67
2	23	55	20	98	9,604	32.67
3	30	65	25	120	11,400	40.00
4	40	45	45	130	16,900	43.33
5	50	80	40	170	28,900	56.67
6	35	75	40	150	22,500	50.00
7	30	63	25	118	13,924	39.33
8	25	35	30	90	8,100	30.00
9	43	75	60	178	31,684	59.33
10	15	58	25	98	9,604	32.67
11	45	80	35	160	25,600	53.33
12	25	65	33	123	15,129	41.00
	$\sum X = 396$	756	408	$\sum X_{\text{subj}} = 1{,}560$	$\sum X^2_{\text{subj}} = 211{,}970$	
	$\sum X^2 = 14{,}228$	49,708	15,214	$\sum X^2_{\text{total}} = 79{,}150$		
	$n = 12$	12	12			
	$\overline{X} = 33$	63	34			$\overline{\overline{X}} = 43.33$

SOURCE	*SS*	*df*	s^2	*F*
Subjects	3,056.03	11	277.82	—
Between groups	6,968.0	2	3,484.0	50.23
Residual	1,525.97	22	69.36	
Total	11,550	35		

at $\alpha = .05$ is 3.44. If the null hypothesis is true, only 5% of the time would we obtain a value of F that equals or exceeds this value. Therefore, we reject H_0 and conclude that different dosages of the new drug have different effects on hyperactive children. As with one-way ANOVA using independently drawn samples, we must now conduct *post hoc* tests to determine which groups differed from the others.

Tukey's HSD test may also be used with ANOVA for repeated measures. (As before, our overall obtained F ratio must first be significant.) However, instead of using s^2_{w}, we use s^2_{resid} to calculate HSD:

Tukey's HSD test critical HSD value for repeated-measures ANOVA = $q\sqrt{\dfrac{s^2_{\text{resid}}}{n}}$

CRITICAL HSD VALUE FOR TUKEY'S
HSD TEST WITH REPEATED MEASURES

$$\text{HSD} = q\sqrt{\frac{s^2_{\text{resid}}}{n}} \tag{18.23}$$

In our example, $s^2_{\text{resid}} = 69.36$ and $n = 12$. To calculate the Studentized range statistic, look in Table F in Appendix D and find the value of q for $df_{\text{resid}} = 11$ and $k = 3$. For $\alpha = .05$, the value of q is 3.82. The critical HSD value is thus

$$\text{HSD} = 3.82\sqrt{\frac{69.36}{12}} = 9.18$$

This is the minimum difference needed between the means of any two groups to conclude that they are significantly different from one another. As you can see from Table 18.6, the moderate dosage of the drug resulted in significantly higher reading scores ($\overline{X} = 63$) for the hyperactive children than were found with either the zero dosage ($\overline{X} = 33$) or the high dosage ($\overline{X} = 34$), which did not significantly differ.

What about effect size? There are several problems with using omega-squared as an estimate of effect size in the repeated-measures design (Keppel, 1991). One measure of effect size that is appropriate for the one-way repeated-measures design is eta-squared:

eta-squared (η^2) an *r* family measure of effect size for repeated-measures ANOVA $\frac{SS_{\text{subj}}}{SS_{\text{subj}} + SS_{\text{resid}}}$

FORMULA FOR ETA-SQUARED

$$\eta^2 = \frac{SS_{\text{subj}}}{SS_{\text{subj}} + SS_{\text{resid}}} \tag{18.24}$$

Eta-squared shows the proportion of variability due to the independent variable after we have removed the variability due to individual differences. In our example (Table 18.6),

$$\eta^2 = \frac{6{,}968}{6{,}968 + 1{,}525.97} = .82$$

assumption of sphericity the assumption that across any two treatment conditions, the variance of the difference scores in the population is the same

Our calculated value of η^2 indicates that 82% of the variability in scores on the reading tests are due to the dosage of the drug. This is a very strong effect.

The assumptions of the one-way ANOVA for repeated measures are that the sample has been independently and randomly drawn, and that for each condition the scores are normally distributed and have an equal variance (see Section 18.11). However, the repeated-measures design has an additional assumption—the **assumption of sphericity**. This is the assumption that across any two conditions, the variance of the difference scores in the population is the same. Substantial departure from these assumptions will affect the accuracy of the ANOVA test.

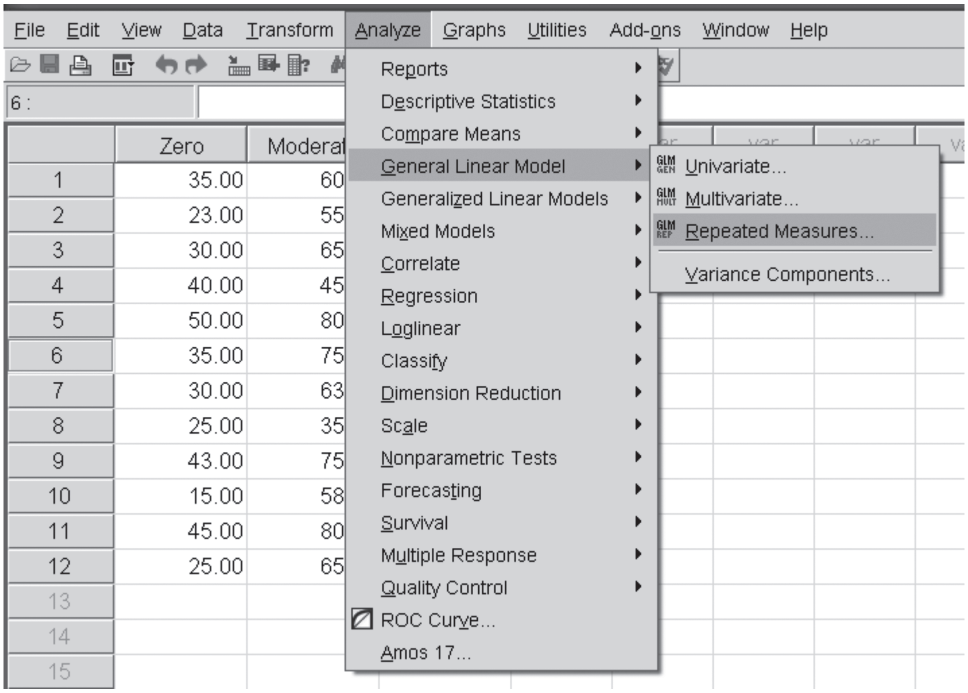
18.19 Calculation of ANOVA for Repeated Measures with IBM SPSS

In this section, you will learn how to use IBM SPSS to calculate the *F* statistic for a one-way ANOVA using repeated measures. You will use the data from Table 18.6.

In the **Data Editor**, all the scores for the dependent variable (i.e., reading scores) should be entered in three different columns—one for each dosage level of the repeated measures. So, in IBM SPSS, because each row will represent a different subject, there will be 12 rows and 3 columns of data. In the **Variable View**, for the three variables, we will use Zero, Moderate, and High, respectively, as the variable names.

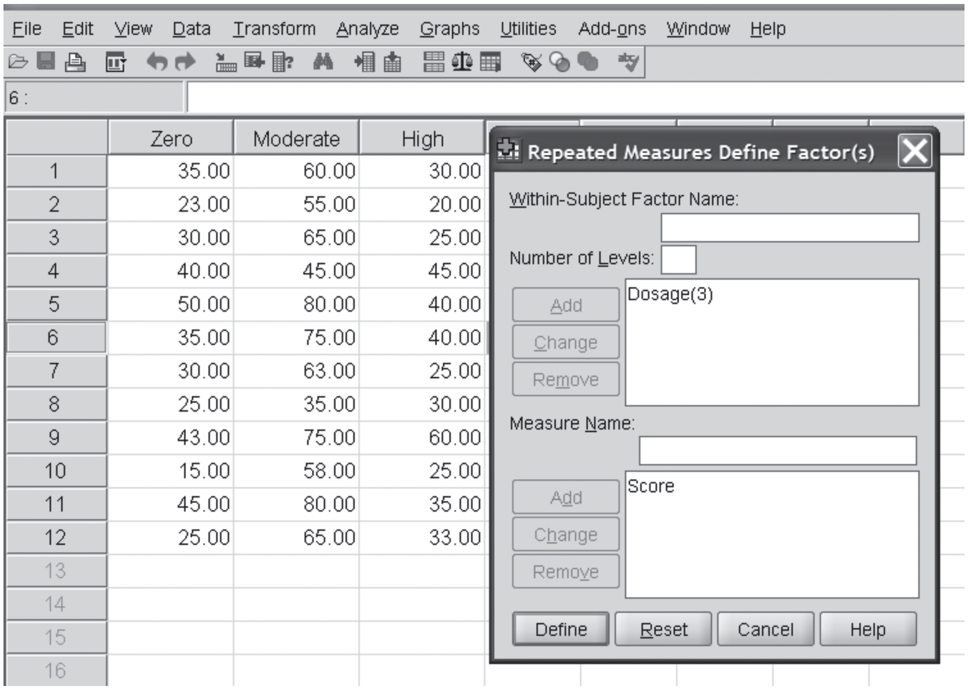
To conduct a hypothesis test for a one-way ANOVA using repeated measures, click **Analyze,** then select **General Linear Model** followed by **Repeated Measures**. Figure 18.10 shows a screenshot of the data as it would appear in the **Data Editor**. In addition, in the figure, you will see the menus that would appear after clicking **Analyze → General Linear Model → Repeated Measures**.

Now, a dialog box will appear. In the box labeled Within-Subject Factor Name, type "Dosage," and for Number of Levels, type "3." Click Add. Then, in the box labeled Measure Name, type "Score." Figure 18.11 shows the dialog box after the labels and levels have been correctly


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
6 :
Reports
Descriptive Statistics
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Quality Control
ROC Curve...
Amos 17...
Univariate...
Multivariate...
Repeated Measures...
Variance Components...

	Zero	Moderat
1	35.00	60
2	23.00	55
3	30.00	65
4	40.00	45
5	50.00	80
6	35.00	75
7	30.00	63
8	25.00	35
9	43.00	75
10	15.00	58
11	45.00	80
12	25.00	65
13		
14		
15		

FIGURE 18.10 IBM SPSS menus when calculating ANOVA for repeated measures.


File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
6 :
Repeated Measures Define Factor(s)
Add
Change
Remove
Add
Change
Remove
Define
Reset
Cancel
Help

	Zero	Moderate	High
1	35.00	60.00	30.00
2	23.00	55.00	20.00
3	30.00	65.00	25.00
4	40.00	45.00	45.00
5	50.00	80.00	40.00
6	35.00	75.00	40.00
7	30.00	63.00	25.00
8	25.00	35.00	30.00
9	43.00	75.00	60.00
10	15.00	58.00	25.00
11	45.00	80.00	35.00
12	25.00	65.00	33.00
13			
14			
15			
16			

Within-Subject Factor Name:

Number of Levels:

Dosage(3)

Measure Name:

Score

FIGURE 18.11 IBM SPSS dialog box when specifying repeated measures in ANOVA.

TABLE 18.7 *IBM SPSS Output from a Repeated Measures ANOVA*

```
GLM Zero Moderate High
  /WSFACTOR=Dosage 3 Polynomial
  /MEASURE=Score
  /METHOD=SSTYPE (3)
  /CRITERIA=ALPHA (.05)
  /WSDESIGN=Dosage.
```

Tests of Within-Subjects Effects

Measure: Score

Source		**Type III Sum of Squares**	***df***	**Mean Square**	***F***	**Sig.**
Dosage	Sphericity Assumed	6968.000	2	3484.000	50.250	.000
	Greenhouse-Geisser	6968.000	1.517	4594.470	50.250	.000
	Huynh-Feldt	6968.000	1.708	4079.214	50.250	.000
	Lower-bound	6968.000	1.000	6968.000	50.250	.000
Error (Dosge)	Sphericity Assumed	1525.333	22	69.333		
	Greenhouse-Geisser	1525.333	16.683	91.432		
	Huynh-Feldt	1525.333	18.790	81.178		
	Lower-bound	1525.333	11.000	138.667		

entered. Click Define. Now, a new dialog box will appear. You will simply move the three variables (Zero, Moderate, and High) to the box labeled Within-Subjects Variables. Click OK.

Because IBM SPSS produces six different tables (much of which is beyond the scope of this book), in Table 18.7, we show only an excerpt of the IBM SPSS output. After the syntax, we present a section of output labeled Tests of Within-Subjects Effects. For the row labeled Dosage, Sphericity Assumed, this corresponds to the row called Between Groups in the ANOVA table at the bottom of Table 18.6. For the row in the IBM SPSS output labeled Error(Dosage), Sphericity Assumed, this corresponds to the row called Residual in the ANOVA table at the bottom of Table 18.6. So, IBM SPSS calculates the *F* ratio assuming that the sphericity assumption has been met. You will notice other rows labeled Greenhouse-Geisser, Huynh-Feldt, and Lower-bound, which are alternative procedures when the sphericity assumption is violated. Both Hays (1994) and Kirk (1995) provide in-depth discussion of the sphericity assumption and the alternative procedures when the assumption is violated.

18.20 Summary

Despite its name, analysis of variance is really a test about means (at least in its simplest forms). One-way ANOVA may be considered an extension of the *t* test to problems involving more than two groups, or conversely, the *t* test may be thought of as a special use of one-way ANOVA.

ANOVA employs the sum of squares (*SS*) as the basic unit of measurement. In one-way ANOVA for independently drawn groups, the total sum of squares (SS_{total}) for all scores (across all groups) is partitioned into two parts: SS_{within}, a measure of the within-group variation of individual scores, and SS_{between}, a measure of the variation of sample means among groups.

To derive variance estimates, we divide *SS*s by their degrees of freedom. When H_0 is true ($\mu_A = \mu_B = \mu_C \ldots$), both types of variance estimate inherent variation, which reflects the variation in behavior of individuals tested under identical condition. Thus, when H_0 is true, s^2_{w} and s^2_{bet} should be equal within the limits of sampling error. When H_0 is false, the within-groups variance is unaffected (because all subjects within a group are still tested identically), but the between-groups variance now reflects the inherent variance *plus* treatment effect. The greater the treatment effect, the greater the value of s^2_{bet}.

ANOVA examines the ratio (called the *F* ratio in honor of Fisher) of these two independent variance estimates. If H_0 is true, $F = s^2_{\mathrm{bet}}/s^2_{\mathrm{w}}$ will approximate unity; if H_0 is false, *F* will be greater than unity (a result of the treatment effect reflected in s^2_{bet}). There is a family of theoretical *F* distributions that depend on the degrees of freedom for both s^2_{bet} and s^2_{w}. As was the case with the *t* test, ANOVA is based on the assumptions of random sampling, normally distributed populations of scores, and homogeneity of variance.

The alternative hypothesis (H_{A}) states only that the group means are unequal in some (any) way. When a significant *F* is obtained, additional *post hoc* tests are employed to determine whether differences exist between specific groups. These tests compensate for the increasing probability, as we make numerous tests, of claiming one or more significant differences when in fact none exist. Among the many possibilities, Tukey's HSD test is a useful *post hoc* test that examines all pairwise comparisons between treatment group means.

As an alternative to ANOVA, we may use planned comparisons. These comparisons, which must be planned in advance of collecting any data, are the most powerful tool available to investigators. Planned comparisons employ a variation of the *t* test, but must be limited to $k - 1$ orthogonal (independent) comparisons.

In analysis of variance for repeated measures, s^2_{w} is further divided into two parts: s^2_{subjects}, a measure of the variability in scores due to individual differences, and s^2_{residual}, a measure of the variability due to random error or chance. The *F* ratio used for this ANOVA is $s^2_{\mathrm{bet}}/s^2_{\mathrm{resid}}$.

Mathematical Note

Note 18.1 Partition of the Sum of Squares and Degrees of Freedom in One-Way Analysis of Variance (*Ref.*: Section 18.3)

$$(X - \overline{\overline{X}}) = (X - \overline{X}) + (\overline{X} - \overline{\overline{X}})$$

Squaring both sides, we obtain

$$(X - \overline{\overline{X}})^2 = (X - \overline{X})^2 + 2(X - \overline{X})(\overline{X} - \overline{\overline{X}}) + (\overline{X} - \overline{\overline{X}})^2$$

Summing within each sample, we obtain

$$\sum^{n_i} (X - \overline{\overline{X}})^2 = \sum^{n_i} (X - \overline{X})^2 + 2(\overline{X} - \overline{\overline{X}}) \sum^{n_i} (X - \overline{X}) + n_i(\overline{X} - \overline{\overline{X}})^2$$

Because $\Sigma(X - \overline{X}) = 0$ (see Note 4.1), the middle term drops out. Then, summing over the *k* subgroups, we obtain

$$\sum^{k}\left[\sum^{n}(X - \overline{\overline{X}})^2\right] = \sum^{k}\left[\sum^{n_i}(X - \overline{X})^2\right] + \sum^{k} n(\overline{X} - \overline{\overline{X}})^2$$

$$\text{or} \qquad SS_{\mathrm{total}} \qquad = \qquad SS_{\mathrm{w}} \qquad + \qquad SS_{\mathrm{bet}}$$

Counting degrees of freedom for each term, we obtain

$$\sum^{k} n - 1 = \sum^{k} (n - 1) + (k - 1)$$

or

$$df_{\text{total}} = df_{\text{w}} + df_{\text{bet}}$$

Proof of the proposition follows from

$$\sum^{k} n - 1 = \sum^{k} n - k + k - 1 = \sum^{k} n - 1$$

Key Terms, Concepts, and Symbols

analysis of variance	(p. 300)
ANOVA	(p. 300)
one-way analysis of variance	(p. 300)
treatment conditions	(p. 300)
levels	(p. 300)
factor	(p. 300)
k	(p. 301)
treatment effect	(p. 301)
within-groups variation	(p. 301)
between-groups variation	(p. 302)
grand mean	(p. 302)
inherent variation	(p. 303)
$\overline{\overline{X}}$	(p. 303)
SS_{total}	(p. 303)
SS_{within}	(p. 304)
SS_{bet}	(p. 304)
partition of the sums of squares	(p. 304)
assumption of homogeneity of variance	(pp. 226, 304)
df_{w}, df_{bet}, df_{total}	(p. 305)
within-groups variance estimate	(p. 305)
s^2_{w}	(p. 305)
between-groups variance estimate	(p. 306)
s^2_{bet}	(p. 306)
F ratio	(p. 306)
F distribution	(p. 306)
t^2	(p. 310)
eta-squared	(pp. 316, 328)
η^2	(pp. 316, 328)
omega-squared	(p. 316)
$\hat{\omega}^2_{\text{bet}}$	(p. 316)
post hoc comparisons	(p. 318)
Tukey's HSD test	(p. 318)
Studentized range statistic	(p. 318)
q	(p. 318)
$\tilde{n}$	(p. 319)
planned comparison	(p. 320)
orthogonal comparison	(p. 321)
ANOVA for repeated measures	(p. 324)
SS_{subjects}	(p. 324)
SS_{resid}	(p. 324)
assumption of sphericity	(p. 328)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In the introduction to the chapter, we used the example of seven experimental drugs being tested for use with Alzheimer's patients. If we use the t test for independent groups to make all pairwise comparisons, we would need to make 21 comparisons. List them. Suppose one of them proved to be significant at the .05 level. Why might the National Institutes of Health not have much faith in our finding?

2. For the example in Problem 1, we decide to use one-way ANOVA to compare the seven group means. (a) Express H_0 in symbolic form. (b) Why can't H_A be expressed in simple symbolic form? (c) Why can't we express H_A as $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6 \neq \mu_7$? (d) List several possible ways in which H_0 can be false.

3. In the analysis of variance *F* test, why is it that only the area in the upper tail of the *F* distribution is of interest?

4. With reference to Table E in Appendix D, what tabulated value of *F* is so great that it would be exceeded in random sampling only 5% of the time when there are (a) 5 degrees of freedom in the numerator and 20 in the denominator? (b) 20 degrees of freedom in the numerator and 5 in the denominator? (c) 3 degrees of freedom in the numerator and 40 in the denominator? (d) 6 degrees of freedom in the numerator and 100 in the denominator?

5. In a one-way ANOVA, equal numbers of subjects were distributed among the several groups. (a) If $df_{\text{bet}} = 3$ and $df_{\text{w}} = 84$, how many groups were there? How many subjects were assigned to each group? What is the numerical value of df_{total}? (b) If $s^2_{\text{bet}} = 170$ and $s^2_{\text{w}} = 50$, was there a significant treatment effect if $\alpha = .05$? If $\alpha = .01$? Justify your answers. (Use the nearest values in Table E to answer this question.)

6. Explain what is meant by (a) inherent variation and (b) treatment effects.

7. We employ one-way ANOVA to compare the performance scores from five groups of nine cases each and obtain a significant *F*. (a) What is the only interpretation that we can make with a significant *F*? (b) What must we do at this point to justify a more specific conclusion? For this analysis, $s^2_{\text{w}} = 18.2$, and the treatment means are as follows: $\overline{X}_1 = 14.5$, $\overline{X}_2 = 11.5$, $\overline{X}_3 = 19.0$, $\overline{X}_4 = 18.1$, and $\overline{X}_5 = 12.0$. (c) Display the difference between the means for all possible pairs of samples as shown in Section 18.14. (d) Apply Tukey's HSD test with $\alpha = .05$. What do you conclude? (e) Apply Tukey's HSD test, but use $\alpha = .01$.

8. Nine chronically hospitalized schizophrenics are randomly assigned to three therapy programs (this would, of course, be far too few subjects for a practical study). At the end of 3 months, each is rated on the "normality" of his or her behavior in a standardized interview situation. The results are shown in Data 18A (higher scores reflect more normal behavior).

DATA 18A

There are three subjects in each of three treatment programs.

	SCORES		
Program *L*:	2	7	3
Program *M*:	6	8	4
Program *N*:	9	11	10

(a) Find SS_{w}, SS_{bet}, and SS_{total} with the deviation-score formula. Does $SS_{\text{total}} = SS_{\text{bet}} + SS_{\text{w}}$? (b) What are the values of df_{w}, df_{bet}, and df_{total}? (c) Find s^2_{w} and s^2_{bet}. (d) Present the outcome developed so far in an ANOVA summary table. (e) Test the hypothesis that $\mu_L = \mu_M = \mu_N$ at $\alpha = .05$, and state your conclusions. (f) If you obtain a significant *F*, apply Tukey's HSD test. (g) Your conclusions? (h) Calculate effect size (eta-squared and omega-squared). (i) Repeat part (a) using the raw-score method.

9. Suppose that the head of the curriculum committee wishes to compare how first-grade students in her school district respond to three new reading programs. At the start of the school year, she randomly selects 15 first-grade students from the district's schools and randomly assigns them to three groups. Each of the three groups will receive one of the reading programs for 9 weeks, and then each student is tested and given a reading score. Due to illness, one of the students in Group B fails to complete the program. The reading scores for the other 14 students are as follows:

DATA 18B

GROUP A	GROUP B	GROUP C
7	12	3
10	10	3
8	14	4
8	11	2
7	13	

(a) Give H_0. (b) Complete the *F* test ($\alpha = .05$) and show your results in a summary table. (c) If you obtain a significant *F*, apply Tukey's HSD test. (d) Your conclusions?

(e) Calculate effect size (eta-squared and omega-squared?)

10. Refer to Problem 9 in Chapter 14 (*t* tests for independent groups). Rework this two-sample problem using one-way ANOVA. Does $F = t^2$?

11. The Always Run Company advertises that its flashlight battery lasts longer than batteries made by other companies. We randomly select five batteries from the latest production run of this company and five batteries each from three of its competitors. We then test their lives in a standard flashlight. The results are given in Data 18C (scores are times in minutes). (a) Give H_0. (b) Complete an *F* test ($\alpha = .05$) and show your results in a summary table. (c) If you obtain a significant *F*, apply Tukey's HSD test. (d) Your conclusions? (e) Calculate effect size (eta-squared and omega-squared).

DATA 18C

ALWAYS RUN	COMPETITOR *A*	COMPETITOR *B*	COMPETITOR *C*
98	94	90	92
92	90	93	86
98	88	92	87
97	88	87	93
94	92	89	92

12. Suppose that you are the president of Always Run Company (Problem 11). You really do not care about all the comparisons. What more powerful test is more appropriate here than one-way ANOVA?

13. (a) If you have four groups, how many comparisons are there in a complete set of independent comparisons? (b) Identify one such set. (c) Identify a different set. (d) Identify another.

14. You wish to compare how a low-fat diet, low-fat diet and a brisk four-mile walk each day, and a low-fat diet and a four-mile jog each day affect cholesterol levels. You randomly assign six subjects to each group. The scores in Data 18D show the reduction in cholesterol count after 2 months. (a) Give H_0. (b) Complete an *F* test ($\alpha = .05$) and show your results in a summary table. (c) Apply Tukey's HSD test. (d) State your conclusions. (e) Calculate effect size.

DATA 18D

LOW-FAT DIET	DIET AND WALK	DIET AND RUN
10	14	19
8	12	15
15	18	14
12	16	16
9	13	18
6	17	20

15. Prior research suggests to you that exercise is the most effective way to reduce cholesterol levels. In place of ANOVA, conduct planned comparisons (that best address your expectation) for the three groups in Problem 14.

16. You wish to compare the short-term effects of three new sleep-aid drugs. You try all three in a sample of eight subjects, with 5 days between each test. Data 18E give the number of hours slept during the night under each drug. (a) Give H_0. (b) Conduct the *F* test ($\alpha = .05$). (c) Apply Tukey's HSD test. (d) State your conclusion. (e) Calculate effect size.

DATA 18E

SUBJECT	DRUG *A*	DRUG *B*	DRUG *C*
1	6	5	8
2	5	6	7
3	6	6	9
4	7	7	6
5	8	6	8
6	4	5	8
7	5	7	8
8	6	6	7

CHAPTER 19

Factorial Analysis of Variance: The Two-Factor Design

When you have finished studying this chapter, you should be able to:

- Explain how analysis of variance can be used to simultaneously study the effects of two variables;
- Understand what is meant by "main effects" and "interaction," and that main effects must be interpreted cautiously when there is a significant interaction;
- Calculate the three F values in a two-factor ANOVA;
- Use IBM SPSS to calculate the three F values in a two-factor ANOVA;
- Make *post hoc* comparisons and calculate effect size; and
- Understand the assumptions for use of the two-factor ANOVA.

Analysis of variance is a broad set of techniques used, for the most part, in studies in which the investigator wishes to compare the means of more than two groups. In the previous chapter, we examined the procedures to use when different groups receive variations of a single treatment variable, such as type of incentive, type of reading program, or different drugs. In

factor
an independent (treatment) variable

this chapter, we move up a step in complexity and examine an analysis of variance design that permits the simultaneous study of the effects of *two* treatment variables or **factors**.

Suppose, for example, that an experimental psychologist wants to know (1) whether the physiological effects of alcohol influence aggression in male social drinkers and (2) whether psychological expectations about alcohol consumption can affect aggression in men. Most of us have witnessed someone's personality change as the result of drinking alcohol. But are these changes due entirely to physiological effects (alcohol is a central nervous system depressant), or can one's expectation of change contribute to his or her becoming less inhibited? The effect on aggression is a particularly important question because alcohol is involved in over half of all violent crimes in the United States.

A psychologist interested in these questions could carry out two separate one-factor experiments, one dealing entirely with the effects of alcohol and the other dealing entirely with expectations about alcohol consumption. If, however, the psychologist were to use the **two-factor analysis of variance design** described in this chapter, he or she would be able to draw conclusions concerning the effects of *both* factors after the completion of a *single* study.

factorial design
an ANOVA design in which two or more factors are studied simultaneously

Lang, Goeckner, Adesso, and Marlatt (1975) pursued these research questions with a clever experimental design that used four groups: (1) expect alcohol–receive alcohol, (2) expect tonic water–receive alcohol, (3) expect alcohol–receive tonic water, and (4) expect tonic water–receive tonic water. (The fourth group served as a control group.) Thus, there were two factors, Alcohol and Expectation, and two conditions for each factor, referred to as **levels** of that factor (alcohol–no alcohol; expect alcohol–do not expect alcohol). This design is called a 2 × 2 (read: 2 by 2) design because there are two levels for each factor and because each level of a factor appears in combination with each level of the other factor. Table 19.1 shows these relationships.

levels
different values of an independent variable (factor)

Lang's team randomly assigned subjects to each of the four groups and misled them into thinking that they were participating in a taste test of different brands of vodka or tonic water. In fact, the experimenters served only two types of drinks, one with alcohol and one without, which pretesting had shown could not be distinguished from one another. Those given alcohol were allowed to drink until they were legally intoxicated (blood-alcohol levels of .10%). After the taste test, the subjects then participated in a "learning study" in which they were the "teacher" and another person (actually, an actor) served as the "student." The subjects were told to administer any of 10 intensities of electric shock whenever the student made a mistake. No current was actually delivered, and the actor pretended to be shocked. The experimenters measured a subject's aggression as the mean chosen level of shock. We present some hypothetical results for this study in Table 19.1.

2 × 2 design
a design with two factors and two levels for each factor; one of many factorial designs (others include 2 × 3, 2 × 4, 3 × 3, etc.)

What you learned in the last chapter has great relevance here. We will partition, or analyze, the total variation into sums of squares associated with different sources of variation, convert them to variance estimates, and form F ratios for the tests of null hypotheses. We will begin our analysis of these results by explaining how the two-factor ANOVA can simultaneously answer the two questions (in our example, one about the effects of alcohol and the other about the expectation of alcohol) that would otherwise require two separate studies.

19.1 Main Effects

In Table 19.1, the two levels of the factor Alcohol are presented in rows and the two levels of the factor Expectation are presented in columns. Each combination of a row level and a column level is referred to as a **cell**. In a 2 × 2 design there are, of course, four cells.

cell
the combination of a row (level of one factor) and a column (level of the other factor)

At the end of the study, we can calculate from the levels of shock administered by each subject during the "learning study," the cell means, the row means, and the column means. Notice in Table 19.1 that the row and column means, shown to the right and at the bottom of the cells, respectively, are equal to the means of the two cell means in the particular row or column. This will always be the case *when there are equal numbers of cases per cell*. Also, because

TABLE 19.1 *Hypothetical Scores in a 2 × 2 Factorial Experiment of the Effects of Alcohol and Expectation of Alcohol on Aggression*

		Expectation		
		Expect alcohol	Do not expect alcohol	
Alcohol	Received alcohol	6 8 5 6 7 8 4 9 6 5 $\sum X = 64$ $\overline{X} = 6.4$	6 3 4 2 4 6 4 3 5 3 $\sum X = 40$ $\overline{X} = 4.0$	$\overline{X}_{\text{alc}} = 5.2$
	Did not receive alcohol	5 7 6 5 7 8 4 4 8 6 $\sum X = 60$ $\overline{X} = 6.0$	5 3 2 4 4 3 3 1 5 4 $\sum X = 34$ $\overline{X} = 3.4$	$\overline{X}_{\text{no alc}} = 4.7$
		$\overline{X}_{\text{expect}} = 6.2$	$\overline{X}_{\text{do not expect}} = 3.7$	$\overline{\overline{X}} = 4.95$

there are equal *n*s, the grand mean ($\overline{\overline{X}}$)—the mean of *all* the scores—is also the mean of the four cell means, the mean of the two row means, and the mean of the two column means.

In this study, we want to know about the *overall* influence of alcohol (in both subjects who expect it and in those who don't) and also about the *overall* influence of the expectation of alcohol (in both subjects who receive alcohol and in those who don't). These are called questions about main effects. The **main effect** for a factor is an "average" effect. It is the differences among the levels of that factor averaged across the levels of the other factor. We can write null hypotheses for each factor as follows:

main effect (for a factor) the mean differences among the levels of one factor averaged across the levels of the other factor

NULL HYPOTHESIS FOR ROW FACTOR (ALCOHOL)

$$H_0\colon \mu_{\text{alc}} = \mu_{\text{no alc}}$$

NULL HYPOTHESIS FOR COLUMN FACTOR (EXPECTATION)

$$H_0\colon \mu_{\text{expect}} = \mu_{\text{do not expect}}$$

When there are only two levels of a factor, as in this case, our alternative hypothesis (H_A) can be either directional (e.g., $\mu_{\text{alc}} > \mu_{\text{no alc}}$) or nondirectional ($\mu_{\text{alc}} \neq \mu_{\text{no alc}}$). When there are more than two levels, the alternative hypothesis is that H_0 is untrue in *some* (any) way and is therefore always nondirectional.

The F test of H_0: $\mu_{\text{alc}} = \mu_{\text{no alc}}$ will tell us whether the overall means for the two alcohol levels ($\overline{X}_{\text{alc}} = 5.2$ and $\overline{X}_{\text{no alc}} = 4.7$) are different enough to justify rejecting the null hypothesis. In essence, we divide our 40 subjects into alcohol and no-alcohol groups of 20 each and perform a variation of one-way ANOVA to compare the overall (row) means of the two groups. Similarly, the F test of H_0: $\mu_{\text{expect}} = \mu_{\text{do no expect}}$ is analogous to a one-way ANOVA to determine whether the overall means of the two levels of expectation ($\overline{X}_{\text{expect}} = 6.2$ and $\overline{X}_{\text{do not expect}} = 3.7$) are significantly different. We use the same 40 subjects, except that we now divide them into two different groups of 20 each, according to their expectations.

Consequently, main effects refer exclusively to differences among the overall row means and differences among the overall column means of tables like 19.1. Notice in Table 19.1, for example, that there is a .5-point overall difference in aggressiveness between subjects who received alcohol and those who did not ($\overline{X}_{\text{alc}} - \overline{X}_{\text{no alc}} = .5$). This is the average of the .4-point difference for the subjects who expected alcohol ($6.4 - 6.0$) and the .6-point difference for the subjects who did not expect alcohol ($4.0 - 3.4$). When interpreting main effects, it is important to keep in mind their "average" nature.

One advantage of using two-way ANOVA over performing a separate experiment for each factor is that conclusions can be broader. For instance, in a separate one-factor experiment to compare the two levels of alcohol consumption, we would hold everything but the Alcohol factor constant, so all the subjects would have to have the same expectations (either of getting alcohol or of getting tonic). Thus, the conclusions regarding alcohol consumption and aggression would apply only to situations characterized by *that* set of expectations. Here we can draw conclusions concerning the effect of alcohol on aggression under two different sets of expectations.

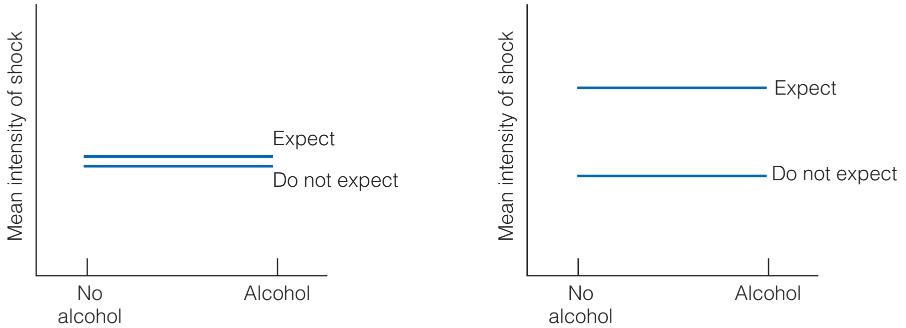
There is another advantage to the factorial design; you will learn about it in the next section.

19.2 Interaction

interaction (between factors) the joint effect of the two independent variables that cannot be predicted from the main effects

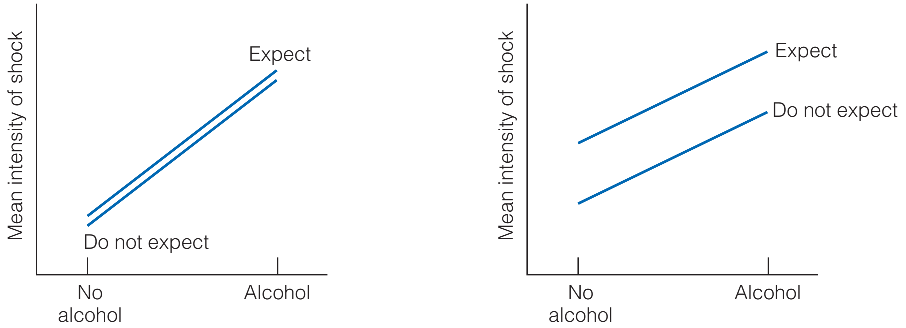
With the factorial design, we are able to examine a new and important question that we cannot examine using the simple t test or the one-way ANOVA: *Whatever the differences among the several levels of one factor, are they the same for each level of the other factor?* In analysis of variance, this kind of question is known as one of interaction effect. **Interaction** refers to the joint effect of the two independent variables (factors) on the dependent variable. We can calculate an F ratio to test for interaction. If the test for interaction is significant, it means that the effect of one factor depends on the level of the other factor.

You can appreciate main effects and interaction effects better by looking at graphs. We present some different hypothetical outcomes of our 2×2 factorial experiment in Figures 19.1*a* through 19.1*f*. In graphing the results of a two-factor experiment, it is customary to display the dependent variable on the vertical axis and the independent variable with the greater number of levels on the horizontal axis. (Thus, in a 2×3 design, the factor with three levels would ordinarily be placed on the horizontal axis, as in Figures 19.1*g* and 19.1*h*.) If, however, one of the independent variables is on a nominal or ordinal scale of measurement and the second is on an interval or ratio scale, it is customary to put the variable on the nominal or ordinal scale along the horizontal axis. In either case, we represent the other independent variable (factor) by lines on the graph that are connected to dots representing the cell means. These are not rigid rules, and each researcher must decide how best to portray his or her data. In our 2×2 example, we have chosen to represent the level of alcohol on the horizontal axis and the expectation of alcohol consumption with lines.



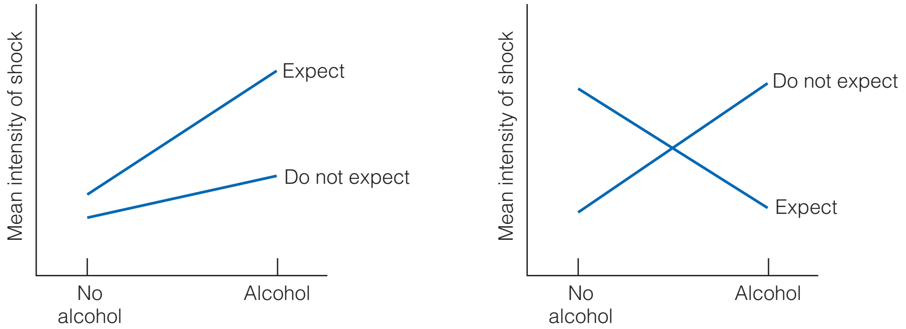
(*a*) No significant effects

(*b*) Significant effect of Expectation; no other effects



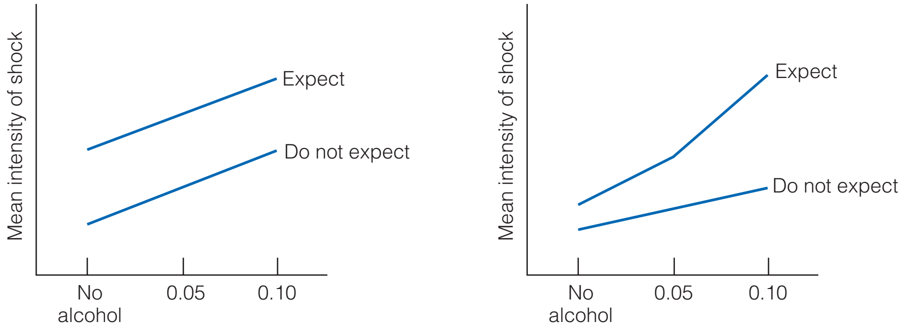
(*c*) Significant effect of Alcohol; no other effects

(*d*) Significant effect of Expectation and of Alcohol; no Interaction effects



(*e*) Significant effect of Expectation, Alcohol, and Interaction

(*f*) Significant Interaction effect; no other effects



(*g*) Significant effect for Expectation and Alcohol; no Interaction effect

(*h*) Significant effect for Expectation, Alcohol, and Interaction

FIGURE 19.1 Possible outcomes of a hypothetical 2 × 2 design (*a–f*) and a hypothetical 2 × 3 design (*g* and *h*) investigating the effects of alcohol consumption and the expectation of alcohol on aggression measured by the intensity of the shock administered.

Figure 19.1*a* shows the results of our experiment as they would appear if there were no main effects and no interaction effect. The mean intensity of shock chosen by the subjects is the same regardless of their expectation of alcohol or their consumption of alcohol.

Figure 19.1*b* shows a main effect for Expectation but none for Alcohol. The subjects who expected alcohol chose higher-intensity shocks than those who did not expect alcohol, and it did not matter whether or not they were consuming alcohol. Figure 19.1*c*, on the other hand, shows a main effect for Alcohol (the subjects who consumed alcohol chose higher levels of shock than those who did not consume alcohol) but none for Expectation. Figure 19.1*d* displays main effects for both Alcohol and Expectation. In each of the examples so far, the lines are parallel, which indicates that there was no interaction—the differences (if any) between the two levels of one factor were the same for each level of the other factor.

We display two possible examples of interaction effects in Figures 19.1*e* and 19.1*f*. When interaction is displayed graphically, it shows a nonchance departure from a parallel relation among the lines connecting the group means. In these results, the difference between the expect-alcohol and do-not-expect-alcohol groups depends on whether or not alcohol was actually consumed. Notice that the lines in the graphs are no longer parallel. In Figure 19.1*e*, there is a negligible difference between the two expectation groups when no alcohol is consumed, but the group that expected alcohol shows considerably more aggression during intoxication. Thus, the effect of expectation depends on whether or not alcohol is actually consumed. This is an example of what is known as an **ordinal interaction**—the relative ranks for the levels of one factor are the same at all levels of the other factor (i.e., the lines do not cross).

ordinal interaction an interaction for which the relative ranks for the levels of one factor are the same at all levels of the other factor

Figure 19.1*f* shows an even more dramatic interaction, called a **disordinal interaction** because the ranks for the levels of one factor differ across levels of the second factor (i.e., the lines cross). There are no main effects (because the *average* effect of expected alcohol is not different from the *average* effect of do not expect alcohol), only an interaction effect.

disordinal interaction an interaction for which the relative ranks for the levels of one factor differ across levels of the other factor

To this point, we have used a 2 × 2 example of a factorial design. Suppose, however, that in addition to the completely sober and intoxicated states, our researchers had included another group with an intermediate level of alcohol consumption (e.g., a 0.05% blood-alcohol level). We show two possible outcomes in Figures 19.1*g* and 19.1*h*. Figure 19.1*g* again shows main effects for both Alcohol and Expectation but no interaction effect. Figure 19.1*h* shows an ordinal interaction effect. The difference between the expect-alcohol and do-not-expect-alcohol groups is not the same for all three levels of alcohol consumption, but the relative ranks are the same. The effect of expectancy on aggression is greater at higher levels of alcohol consumption.

The third null hypothesis to be tested, then, in a two-factor ANOVA is that of *no interaction*. It can be stated in general terms as follows: *When there is no interaction, the population differences in performance among the several levels of one factor are the same regardless of the level of the other factor.* If in the samples the differences among the levels of one factor fluctuate more from level to level of the other factor than would be expected on the basis of sampling variation, we reject the null hypothesis and conclude that the effects of one factor do in fact depend on the particular level of the other factor in the population.

19.3 The Importance of Interaction

Because a main effect is an *average effect* (i.e., it reflects the *average* difference across the levels of the other treatment variable), it can be misleading when an interaction is present. Remember Minium's First Law of Statistics: "The eyeball is the statistician's most powerful instrument." Look again at Figure 19.1*g*. In this case, we would make the general conclusion that the expect-alcohol subjects are more aggressive than the do-not-expect-alcohol subjects—it applies equally well at all levels of the second factor. However, if the results were those of Figure 19.1*h* (in which there is an ordinal interaction), the same conclusion about the main effect would be misleading. It fails to take note of the fact that the difference between the expect-alcohol and do-not-expect-alcohol subjects is negligible for the no-alcohol condition, whereas the difference is large when considerable alcohol has been consumed. We must be very careful when reporting main effects if there is a disordinal interaction. In this case, we cannot interpret the main effect without referring to the interaction.

The presence of a significant interaction should therefore be a red flag to us when it comes to interpretation. *When interaction is present we should examine the effects of any factor of interest at each level of the interacting factor before making interpretations*. Certainly the best way to begin is to consult the table of cell means and a graph of these values. Remember that a two-factor design is really made up of several one-factor experiments. In addition to main effects, the factorial design also allows us to test for what are called *simple effects*. Consider, for example, a 3 × 3 design. If we refer to one factor as A and the other as B, then the three levels of each are A_1, A_2, and A_3 and B_1, B_2, and B_3, respectively. Main effects compare differences among the levels of one factor *averaged* across all levels of the other. However, this particular design consists of six one-way experiments, and if we wish, we may analyze each of them separately. Perhaps we are interested in the effects of A (all three levels) specifically for condition B_2. **Simple effects** refer to the results of these one-factor analyses. To make such comparisons, the interaction must first be significant and if it is, we may use procedures such as Tukey's HSD test (Section 18.14).

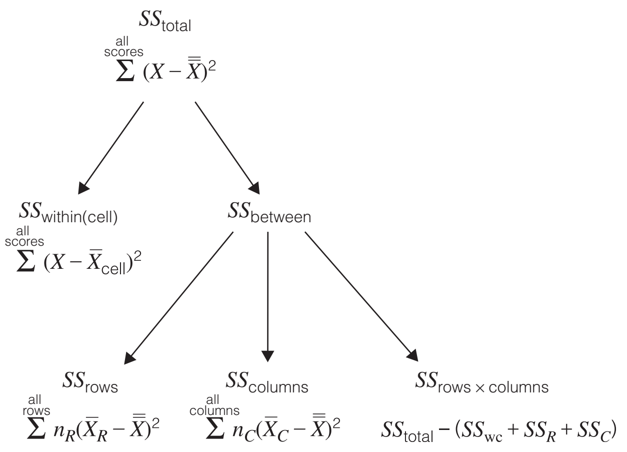
simple effects results of single-factor analyses (differences within a row or a column)

Sometimes the question of interaction is substantially more important than that of a main effect. Consider, for example, a study involving two methods of teaching and learners of low or high aptitude. Researchers long ago established that individuals with low aptitude learn more slowly than individuals with high aptitude; we really are not interested in confirming that effect. The real point of the study is to explore whether the relative difference in effectiveness of the two teaching methods is the same for slow learners as for fast ones, a question about the interaction between method and learning ability. It could well be, for instance, that method and learning ability interact to such a degree that one method is more effective with slow learners and the other is more effective with fast learners. A graphic display of this result would look like Figure 19.1*f*. We will explain how to interpret interaction in greater detail in Section 19.7.

19.4 Partition of the Sums of Squares for Two-Way ANOVA

Now let us begin the step-by-step development of two-way ANOVA for independent groups. We designate one factor the column factor (Expectation in our example, Table 19.1) and designate the other factor the row factor (Alcohol in our example). The two levels of each factor are referred to as C_1 and C_2 and as R_1 and R_2, respectively.

Recall that in one-way ANOVA we partition the total sum of squares into two parts: $SS_{\text{total}} = SS_{\text{w}} + SS_{\text{bet}}$. *In the two-way design,* SS_{bet} *is further partitioned into the three parts shown in Figure 19.2*: $SS_{\text{bet}} = SS_{\text{rows}} + SS_{\text{columns}} + SS_{\text{rows}\times\text{columns}}$ (read: rows by columns). This is because in the factorial design, we consider the rows and columns as groups. Of the five sums of squares to be considered, you should be able to understand the definitional (deviation-score) formulas for the first four from our earlier development of one-way ANOVA. In two-way ANOVA, SS_{rows} reflects the variability due to one factor; SS_{columns} reflects the variability due to the second factor; and $SS_{\text{rows}\times\text{columns}}$ reflects the interaction variability.



SS_{total}
$\sum^{\text{all scores}} (X - \bar{\bar{X}})^2$
$SS_{\text{within(cell)}}$
$\sum^{\text{all scores}} (X - \bar{X}_{\text{cell}})^2$
SS_{between}
SS_{rows}
$\sum^{\text{all rows}} n_R(\bar{X}_R - \bar{\bar{X}})^2$
SS_{columns}
$\sum^{\text{all columns}} n_C(\bar{X}_C - \bar{\bar{X}})^2$
$SS_{\text{rows} \times \text{columns}}$
$SS_{\text{total}} - (SS_{\text{wc}} + SS_R + SS_C)$

FIGURE 19.2 Partitioning of the sums of squares for a two-way ANOVA for independent groups.

Although computer packages for factorial ANOVA designs are widely available, we will present the raw-score formulas as well. The numbers of cases are distinguished by subscript: n_{total} is the total number of cases, n_R is the number of cases per row, n_C is the number of cases per column, and n_{cell} is the number of cases per cell.

SS_{total} sum of the squared deviations from the grand mean; $\sum^{\text{all scores}} (X - \bar{\bar{X}})^2$

1. *Total sum of squares:* **SS_{total}**. This is identical to SS_{total} in the one-way design. It is a measure of total variation present in our data and is generated from the squared deviations of the individual scores from the grand mean of all scores:

TOTAL SUM OF SQUARES:
DEVIATION-SCORE FORMULA

$$SS_{\text{total}} = \sum^{\text{all scores}} (X - \bar{\bar{X}})^2 \qquad (19.1a)$$

TOTAL SUM OF SQUARES:
RAW-SCORE FORMULA

$$SS_{\text{total}} = \sum^{\text{all scores}} X^2 - \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}} \qquad (19.1b)$$

SS_{wc} sum of the squared deviations within cells (analogous to SS_{within} in the single-factor ANOVA), $\sum^{\text{all scores}} (X - \bar{X}_{\text{cell}})^2$

2. *Within-cells sum of squares:* **SS_{wc}** (sometimes called SS_{error}). This would be the same as SS_{w} (within groups) in the one-way design if each cell were considered a group. It is based on the squared deviations of the individual scores from their cell means:

WITHIN-CELLS SUM OF SQUARES:
DEVIATION-SCORE FORMULA

$$SS_{\text{wc}} = \sum^{\text{all scores}} (X - \bar{X}_{\text{cell}})^2 \qquad (19.2a)$$

WITHIN-CELLS SUM OF SQUARES:
RAW-SCORE FORMULA

$$SS_{\text{wc}} = \sum^{\substack{\text{all}\\ \text{cells}}} \left[\sum X^2 - \frac{\left(\sum X\right)^2}{n_{\text{cell}}} \right] \tag{19.2b}$$

or

WHEN EQUAL *n* PER CELL

$$SS_{\text{wc}} = \sum^{\substack{\text{all}\\ \text{scores}}} X^2 - \frac{\sum^{\substack{\text{all}\\ \text{cells}}} \left(\sum^{\text{cell}} X \right)^2}{n_{\text{cell}}} \tag{19.2c}$$

SS_R
sum of the squared deviations for rows (one factor); $\sum^{\substack{\text{all}\\ \text{rows}}} n_R(\overline{X}_R - \overline{\overline{X}})^2$

3. *Row sum of squares:* **SS_R**. This is analogous to SS_{bet} in the one-way design where each row is considered a group. It is based on the squared deviations of the row means from the grand mean:

ROW SUM OF SQUARES: DEVIATION-SCORE FORMULA

$$SS_R = \sum^{\substack{\text{all}\\ \text{rows}}} n_R(\overline{X}_R - \overline{\overline{X}})^2 \tag{19.3a}$$

ROW SUM OF SQUARES: RAW-SCORE FORMULA

$$SS_R = \left[\frac{\left(\sum X_{R_1}\right)^2}{n_{R_1}} + \frac{\left(\sum X_{R_2}\right)^2}{n_{R_2}} + \cdots \right] - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X \right)^2}{n_{\text{total}}} \tag{19.3b}$$

or

WHEN EQUAL *n* PER ROW

$$SS_R = \left[\frac{\left(\sum X_{R_1}\right)^2 + \left(\sum X_{R_2}\right)^2 + \cdots}{n_R} \right] - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X \right)^2}{n_{\text{total}}} \tag{19.3c}$$

SS_C
sum of the squared deviations for columns (other factor) $\sum^{\substack{\text{all}\\ \text{columns}}} n_C(\overline{X}_C - \overline{\overline{X}})^2$

4. *Column sum of squares:* **SS_C**. The analogy to SS_{bet} holds here as well, except now the columns become the groups. SS_C is generated from the squared deviations of the column means from the grand mean:

COLUMN SUM OF SQUARES: DEVIATION-SCORE FORMULA

$$SS_C = \sum^{\substack{\text{all}\\ \text{columns}}} n_C\,(\overline{X}_C - \overline{\overline{X}})^2 \tag{19.4a}$$

COLUMN SUM OF SQUARES: RAW-SCORE FORMULA

$$SS_C = \left[\frac{\left(\sum X_{C_1}\right)^2}{n_{C_1}} + \frac{\left(\sum X_{C_2}\right)^2}{n_{C_2}} + \cdots\right] - \frac{\left(\sum\limits^{\text{all scores}} X\right)^2}{n_{\text{total}}} \qquad (19.4b)$$

or

WHEN EQUAL *n* PER COLUMN

$$SS_C = \left[\frac{\left(\sum X_{C_1}\right)^2 + \left(\sum X_{C_2}\right)^2 + \cdots}{n_C}\right] - \frac{\left(\sum\limits^{\text{all scores}} X\right)^2}{n_{\text{total}}} \qquad (19.4c)$$

$SS_{R\times C}$ interaction sum of squares; $SS_{\text{total}} - (SS_{\text{wc}} + SS_R + SS_C)$

5. *Interaction sum of squares:* **$SS_{R\times C}$** (read: row by column sum of squares). This is the new element. $SS_{R\times C}$ is generated from the squared deviations of the cell means from what would be predicted if there were no interaction. Because the several components (SS_{wc}, SS_R, SS_C, and $SS_{R\times C}$) sum to SS_{total}, we may find $SS_{R\times C}$ most easily by subtraction:

INTERACTION SUM OF SQUARES

$$SS_{R\times C} = SS_{\text{total}} - (SS_{\text{wc}} + SS_R + SS_C) \qquad (19.5)$$

Be sure to check your calculations; an error anywhere along the line will also cause an error in $SS_{R\times C}$. Remember also that a negative sum of squares is impossible; it results from an error in calculation.

As an example, refer again to Table 19.1. We will calculate the various sums of squares in our Alcohol × Expectation example. We start, as in one-way ANOVA, by obtaining ΣX and ΣX^2 for each group—in this case, for each cell. These are the basic building blocks from which all other computations flow.

	C_1	C_2
R_1	$\sum X = 6 + 8 + \cdots + 5 = 64$ $\sum X^2 = 36 + 64 + \cdots + 25 = 432$	$\sum X = 6 + 3 + \cdots + 3 = 40$ $\sum X^2 = 36 + 9 + \cdots + 9 = 176$
R_2	$\sum X = 5 + 7 + \cdots + 6 = 60$ $\sum X^2 = 25 + 49 + \cdots + 36 = 380$	$\sum X = 5 + 3 + \cdots + 4 = 34$ $\sum X^2 = 25 + 9 + \cdots + 16 = 130$

Although we would not use a table like this in practice, it *is* a good idea to record the values of ΣX and ΣX^2 obtained in an appropriate cell diagram of the two-way design, as shown in the following table:

		C_1	C_2	
$n_{\text{cell}} = 10$	R_1	$\sum X = 64$	$\sum X = 40$	$\sum X_{R_1} = 104$
$n_R = 20$		$\sum X^2 = 432$	$\sum X^2 = 176$	
$n_C = 20$	R_2	$\sum X = 60$	$\sum X = 34$	$\sum X_{R_2} = 94$
		$\sum X^2 = 380$	$\sum X^2 = 130$	
$n_{\text{total}} = 40$		$\sum X_{C_1} = 124$	$\sum X_{C_2} = 74$	$\sum\limits^{\substack{\text{all}\\\text{scores}}} X = 198$
				$\sum\limits^{\substack{\text{all}\\\text{scores}}} X^2 = 1{,}118$

From the diagram, it is a simple matter to obtain the necessary quantities for use in the raw-score formulas. First, for row 1 we add values of ΣX across the row to obtain $\Sigma X_{R_1} = 104$; we then do the same for the remaining row(s). Second, for column 1 we add the values of ΣX down through the rows to obtain $\Sigma X_{C_1} = 124$; we then do the same for the remaining column(s). Third, we add the values of ΣX and ΣX^2 for all cells to obtain

$$\sum\limits^{\substack{\text{all}\\\text{scores}}} X = 198 \text{ and } \sum\limits^{\substack{\text{all}\\\text{scores}}} X^2 = 1{,}118$$

Now we have the appropriate values to substitute in the raw-score formulas:

$$SS_{\text{total}} = \sum\limits^{\substack{\text{all}\\\text{scores}}} X^2 - \frac{\left(\sum\limits^{\substack{\text{all}\\\text{scores}}} X\right)^2}{n_{\text{total}}} = 1{,}118 - \frac{198^2}{40} = 137.9$$

$$SS_{\text{wc}} = \sum\limits^{\substack{\text{all}\\\text{scores}}} X^2 - \frac{\sum\limits^{\substack{\text{all}\\\text{cells}}}\left(\sum\limits^{\text{cell}} X\right)^2}{n_{\text{cell}}} = 1{,}118 - \frac{64^2 + 40^2 + 60^2 + 34^2}{10} = 72.8$$

$$SS_R = \left[\frac{\left(\sum X_{R_1}\right)^2 + \left(\sum X_{R_2}\right)^2}{n_R}\right] - \frac{\left(\sum\limits^{\substack{\text{all}\\\text{scores}}} X\right)^2}{n_{\text{total}}} = \frac{104^2 + 94^2}{20} - \frac{198^2}{40} = 2.5$$

$$SS_C = \left[\frac{\left(\sum X_{C_1}\right)^2 + \left(\sum X_{C_2}\right)^2}{n_C}\right] - \frac{\left(\sum\limits^{\substack{\text{all}\\\text{scores}}} X\right)^2}{n_{\text{total}}} = \frac{124^2 + 74^2}{20} - \frac{198^2}{40} = 62.5$$

$$\begin{aligned} SS_{R\times C} &= SS_{\text{total}} - (SS_{\text{wc}} + SS_R + SS_C) \\ &= 137.9 - (72.8 + 2.50 + 62.5) \\ &= .1 \end{aligned}$$

19.5 Degrees of Freedom

$\boldsymbol{df_{\text{total}}}$
$n_{\text{total}} - 1$

$\boldsymbol{df_R}$
$R - 1$

$\boldsymbol{df_C}$
$C - 1$

$\boldsymbol{df_{R \times C}}$
$(R - 1)(C - 1)$

$\boldsymbol{df_{\text{wc}}}$
$\sum^{\substack{\text{all}\\\text{cells}}}(n_{\text{wc}} - 1)$

Just as with one-way ANOVA, the total degrees of freedom is divided into components associated with each sum of squares:

$$\boldsymbol{df_{\text{total}} = df_R + df_C + df_{R \times C} + df_{\text{wc}}}$$

As before, $df_{\text{total}} = n_{\text{total}} - 1$. In counting degrees of freedom for the other sources of variation, we will let C equal the number of columns, R equal the number of rows, and n_{wc} equal the number of scores within each cell. Because there are C deviations involved in the computation of SS_C, $df_C = C - 1$; similarly, $df_R = R - 1$. In computing SS_{wc}, we consider the deviation of each score in the cell from the cell mean. Consequently, each cell contributes $n_{\text{wc}} - 1$ degrees of freedom, and

$$df_{\text{wc}} = \sum^{\substack{\text{all}\\\text{cells}}}(n_{\text{wc}} - 1)$$

Finally, $df_{R \times C} = (R - 1)(C - 1)$.

In our example (Table 19.1), we calculate the degrees of freedom as follows:

$$\begin{aligned}
df_{\text{total}} &= 40 - 1 = 39\\
df_R &= 2 - 1 = 1\\
df_C &= 2 - 1 = 1\\
df_{R \times C} &= (2 - 1)(2 - 1) = 1\\
df_{\text{wc}} &= (10 - 1) + (10 - 1) + (10 - 1) + (10 - 1) = 36
\end{aligned}$$

19.6 Variance Estimates and *F* Tests

$\boldsymbol{s^2_{\text{wc}}}$
the within-cells variance estimate; an estimate of inherent variance only;
$\dfrac{SS_{\text{wc}}}{df_{\text{wc}}}$

$\boldsymbol{s^2_R}$
the rows variance estimate; an estimate of inherent variance plus main effect for the row factor;
$\dfrac{SS_R}{df_R}$

To calculate an unbiased estimate of the population variance, you recall, we must divide the sum of squares by its degrees of freedom (Section 12.9). We can calculate four variance estimates by dividing SS_{wc}, SS_R, SS_C, and $SS_{R \times C}$ by their respective degrees of freedom.

$$\boldsymbol{s^2_{\text{wc}}} = \frac{SS_{\text{wc}}}{df_{\text{wc}}} \xrightarrow{\text{estimates}} \textit{inherent variance } (\sigma^2)$$

As in the single-factor analysis, the within-cells value is an estimate of inherent variance (σ^2) only and is sometimes called the *error variance.* It is unaffected by row, column, or interaction effects. The inherent variation is assumed to be equal across all populations (cells)—see Section 19.11.

$$\boldsymbol{s^2_R} = \frac{SS_R}{df_R} \xrightarrow{\text{estimates}} \textit{inherent variance + main effect for the row factor}$$

If $\mu_{R_1} = \mu_{R_2} = \ldots$, then the variation among the row means ($\overline{X}_{R_1}, \overline{X}_{R_2}, \ldots$) will reflect only inherent variation. Under these circumstances, s^2_R will estimate the same quantity estimated by s^2_{wc}. If there is a difference among the levels of the row factor, s^2_R will tend to be larger than s^2_{wc}. It is therefore analogous to s^2_{bet} in one-way ANOVA.

$$\boldsymbol{s^2_C} = \frac{SS_C}{df_C} \xrightarrow{\text{estimates}} \textit{inherent variance + main effect for the column factor}$$

TABLE 19.2 *General Form and Example of the Two-Factor ANOVA Summary Table*

SOURCE	*SS*	*df*	s^2	F_{calc}
Rows	Formulas 19.3*a*, 19.3*b*, or 19.3*c*	$R - 1$	$\frac{SS_R}{df_R}$	$\frac{s_R^2}{s_{wc}^2}$
Columns	Formulas 19.4*a*, 19.4*b*, or 19.4*c*	$C - 1$	$\frac{SS_C}{df_C}$	$\frac{s_C^2}{s_{wc}^2}$
Rows × Columns	Formula 19.5	$(R - 1)(C - 1)$	$\frac{SS_{R\times C}}{df_{R\times C}}$	$\frac{s_{R\times C}^2}{s_{wc}^2}$
Within cells	Formula 19.2*a*, 19.2*b*, or 19.2*c*	$\sum^{\text{all cells}} (n_{wc} - 1)$	$\frac{SS_{wc}}{df_{wc}}$	
Total	$SS_{wc} + SS_R + SS_C + SS_{R\times C}$	$n_{\text{total}} - 1$		

SOURCE	*SS*	*df*	s^2	F_{calc}
Rows	2.5	1	2.5	1.24
Columns	62.5	1	62.5	30.94
Rows × Columns	0.1	1	.1	.05
Within cells	72.8	36	2.00	
Total	137.9	39		

s_C^2
the columns variance estimate; an estimate of inherent variance plus main effect for the column factor; $\frac{SS_C}{df_C}$

$s_{R\times C}^2$
the interaction variance estimate; an estimate of inherent variance plus interaction effect; $\frac{SS_{R\times C}}{df_{R\times C}}$

***F* ratios**
$\frac{s_R^2}{s_{wc}^2}, \frac{s_C^2}{s_{wc}^2}, \frac{s_{R\times C}^2}{s_{wc}^2};$
$\frac{\textit{inherent variance} + \textit{treatment effect}}{\textit{inherent variance,}}$
When there is no treatment effect, the ratio should approximate 1.0.

If the independent variable represented in the columns has no effect, then $\mu_{C_1} = \mu_{C_2} = \ldots,$ and the variation among the column means $(\overline{X}_{C_1}, \overline{X}_{C_2}, \ldots)$ will reflect only inherent variation. It is therefore just like s_R^2, except that it is sensitive to column effect rather than to row effect.

$$s_{R\times C}^2 = \frac{SS_{R\times C}}{df_{R\times C}} \xrightarrow{\text{estimates}} \textit{inherent variance} + \textit{interaction effect}$$

If there is no interaction in the populations, $s_{R\times C}^2$ will be responsive only to inherent variation and will estimate the same quantity estimated by s_{wc}^2. If interaction is present, $s_{R\times C}^2$ will respond to it and will therefore tend to be larger.

In one-way ANOVA, we created an *F* ratio by dividing s_{bet}^2 by s_w^2, each an estimate of inherent variation if the null hypothesis is true. Here, we have three *F* ratios calculated by dividing s_R^2, s_C^2, and $s_{R\times C}^2$ by s_{wc}^2. As before, we organize our calculations into a *summary table,* as shown in Table 19.2. We obtain the critical values of *F* from Table E of Appendix D to learn what magnitude we must reach or exceed to declare significance.

In our example, all three *F* ratios have 1 *df* associated with the numerator and 36 *df* associated with the denominator. Thus, the critical value of *F* is the same for all three, 4.11 ($\alpha = .05$). Only the obtained value of *F* for the column effect exceeds its critical value; consequently, we reject the null hypothesis that $\mu_{C_1} = \mu_{C_2}$. These results do, in fact, mimic those found by Lang's research team in their study.

19.7 Studying the Outcome of Two-Factor Analysis of Variance

It is ironic that the two forms of ANOVA that we have studied are fundamentally tests about means and their relationships, yet it is possible to complete the *F* tests without ever calculating a single mean (and some research reports actually do omit this computation). Calculation of cell and marginal means ought to be a routine part of ANOVA. If any of the three F tests is significant, for example, you should always organize the cell means in table or group form. Always look carefully at the means and their relationship to one another to see whether the results make sense. If they do not, the first step is to hunt for possible computational error.

For the present problem, a graph of the four cell means would look almost exactly like Figure 19.1*b*. This figure shows that the male social drinkers who expected alcohol gave higher intensity shocks than did those who did not expect alcohol, regardless of actual alcohol consumption. There was no significant effect for Alcohol, and the nonsignificant interaction component is reflected by the fact that the two lines are reasonably parallel. Thus the main effect of Expectation is generalizable to both levels of alcohol consumption used in the experiment.

If one or both of our main effects prove to be significant, we may then use Tukey's HSD test (or another *post hoc* test) to determine which levels of a factor differ from others:

Tukey's HSD test a commonly used *post hoc* test; critical HSD value = $q\sqrt{\frac{s^2_{\text{wc}}}{n}}$

q Studentized range statistic

TUKEY'S HSD TEST FOR FACTORIAL DESIGNS

$$\text{HSD} = q\sqrt{\frac{s^2_{\text{wc}}}{n}} \qquad (19.6)$$

where: n = the number of cases per row or per column
q = depends on the number of levels for the factor being tested.

Of course, we would use a *post hoc* test only if there are more than two levels. If we find a significant main effect for a factor with only two levels, as in our example, we already know that the two levels are different.

When the interaction is significant, we must also analyze it. This is not as easy as many people think. In fact, a survey of published research articles found that 99% of investigators who employed factorial designs had not interpreted their interaction effect in an unequivocally correct manner (Rosnow & Rosenthal, 1989a). To interpret interactions correctly, we must do more than just look at original cell means because these reflect main effects as well as interaction. We must instead look at the "corrected" cell means—the cell means *after* the main effects have been removed (Rosnow & Rosenthal, 1991, 1995).

As an example, consider the following four original cell means of a 2 × 2 design:

		Factor *B*		Row
		B_1	B_2	means
Factor *A*	A_1	5	5	5
	A_2	7	9	8
	Column means	6	7	$\overline{\overline{X}} = 6.5$

If we were to interpret the interaction by looking only at the differences among the original cell means, as many researchers do, we would wind up with a rank ordering of $A_2B_2 > A_2B_1 > A_1B_1 = A_1B_2$. However, to look at the interaction only, we must first subtract the other influences:

CORRECTED CELL MEANS FOR
INTERPRETING INTERACTION

$$\textit{cell mean} - (\textit{row effect} + \textit{column effect} + \overline{\overline{X}}) \qquad (19.7)$$

where: *row effect* = the row mean minus the grand mean ($\overline{\overline{X}}$)
column effect = the column mean minus the grand mean.

Let us now make the corrections for the cell means:

$$A_1B_1\text{: } 5 - [(5 - 6.5) + (6 - 6.5) + 6.5] = +0.5$$

$$A_1B_2\text{: } 5 - [(5 - 6.5) + (7 - 6.5) + 6.5] = -0.5$$

$$A_2B_1\text{: } 7 - [(8 - 6.5) + (6 - 6.5) + 6.5] = -0.5$$

$$A_2B_2\text{: } 9 - [(8 - 6.5) + (7 - 6.5) + 6.5] = +0.5$$

In this example, a graphic representation of just the interaction effect is X-shaped, but the effect is small. If the main effects were large, the effect of one factor would vary only slightly across the two levels of the other.

19.8 Effect Size

Just as with t or one-way ANOVA, we can (and should) measure the magnitude of experimental effect. First, review Section 18.12. Recall that two measures of effect size used with one-way ANOVA are eta-squared (η^2) and omega-squared ($\hat{\omega}^2$). These two measures are generally used with the factorial design as well.

eta-squared (η^2)
an r family measure of effect size

Eta-squared is the easier of the two—all we do is divide the sum of squares by the total sum of squares. This provides us with the proportion of variance in the dependent variable accounted for by the main and interaction effects:

FORMULAS FOR ETA-SQUARED

$$\eta^2_{\text{rows}} = \frac{SS_R}{SS_{\text{total}}}$$

$$\eta^2_{\text{columns}} = \frac{SS_C}{SS_{\text{total}}} \qquad (19.8)$$

$$\eta^2_{R\times C} = \frac{SS_{R\times C}}{SS_{\text{total}}}$$

In our example (Table 19.2), $\eta^2_R = .018$, $\eta^2_C = .453$, and $\eta^2_{R\times C} = .001$. Thus, there was a strong effect for columns (Expectations) and weak effects for rows (Alcohol) and interaction.

partial omega-squared ($\hat{\omega}^2$)
an r family measure of effect size; the degree of association between the dependent variable and the experimental effect for each factor

Eta-squared is frequently used, and it is easy to calculate, but recall from Section 18.12 that this measure is also biased. A better measure of effect size is called **partial omega-squared** (Keren & Lewis, 1979). Effect size as calculated by $\hat{\omega}^2$ is not influenced by the other experimental design effects. The general formula for $\hat{\omega}^2$ is as follows:

$$\hat{\omega}^2 = \frac{\textit{Estimated variance} \text{ (For } R,\ C, \text{ or } R \times C)}{\textit{Estimated variance} \text{ (for } R,\ C, \text{ or } R \times C) + \textit{Estimated variance within cells}}$$

The computational formulas are as follows:

COMPUTATIONAL FORMULAS FOR
PARTIAL OMEGA-SQUARED

$$\hat{\omega}^2_R = \frac{(R - 1)(F_R - 1)}{(R - 1)(F_R - 1) + n_{\text{cell}}\, RC}$$

$$\hat{\omega}_C^2 = \frac{(C-1)(F_C-1)}{(C-1)(F_C-1)+n_{\text{cell}}RC} \tag{19.9}$$

$$\hat{\omega}_{R\times C}^2 = \frac{(R-1)(C-1)(F_{R\times C}-1)}{(R-1)(C-1)(F_{R\times C}-1)+n_{\text{cell}}RC}$$

where: R = number of rows
C = number of columns
F_R = value of F for rows
F_C = value of F for columns
$F_{R\times C}$ = value of F for interaction.

Partial omega-squared provides us with the degree of association between the dependent variable and the experimental effect for each factor. In our example (Table 19.2), $\hat{\omega}_R^2 = .006$, $\hat{\omega}_C^2 = .428$, and $\hat{\omega}_{R\times C}^2 = .000$.

19.9 Calculation of Two-Factor ANOVA with IBM SPSS

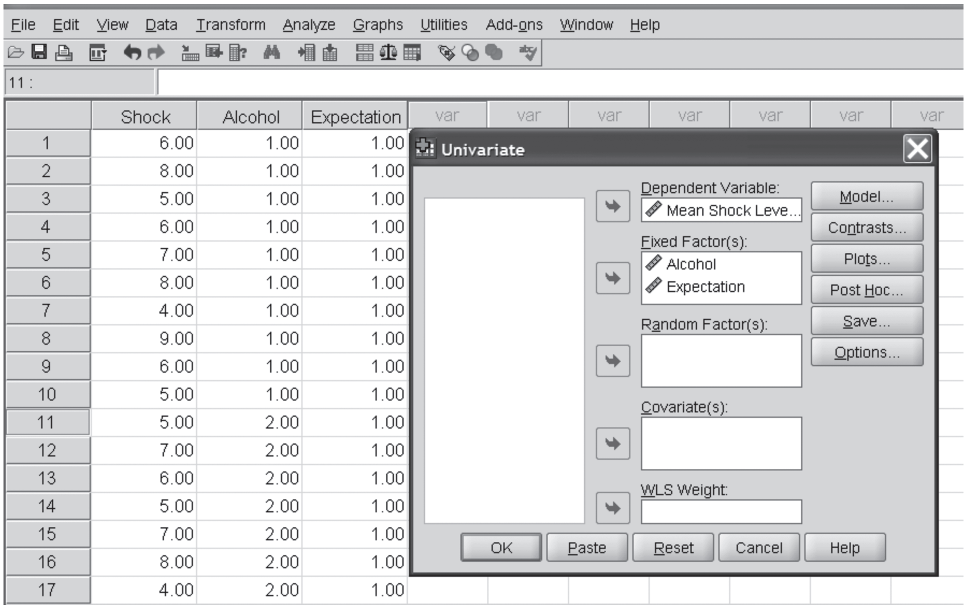
In this section, you will learn how to use IBM SPSS to calculate the three F ratios to test hypotheses about main effects and the interaction effect in a two-way ANOVA. You will use the data from Table 19.1.

In the **Data Editor**, all the scores for the dependent variable (i.e., mean shock level chosen) should be entered in a single column. Now, in the second column, we will identify what level of the row factor the score came from. For our example, we will enter 1s for scores from subjects who received alcohol and 2s for scores from subjects who did not receive alcohol. Now, in the third column, we will identify what level of the column factor the score came from. We will enter 1s for scores from subjects who expected alcohol and 2s for scores from subjects who did not expect alcohol. In the **Variable View**, for our dependent variable, we will use Shock as the variable name and Mean Shock Level Chosen as the label. For the second and third variables, we will simply use Alcohol and Expectation, respectively, for the variable names. Similar to Section 14.7 and 18.10, we will use the column labeled Values. For the Alcohol variable, click on Values. In the Value box, type “1” and, in the Label box, type “Received alcohol.” Click Add. Now, in the Value box, type “2” and, in the Label box, type “Did not receive alcohol.” Click Add. For the Expectation variable, click on Values. In the Value box, type “1” and, in the Label box, type “Expect alcohol.” Click Add. Now, in the Value box, type “2” and, in the Label box, type “Do not expect alcohol.” Click Add, then click OK.

To conduct hypothesis tests for a two-way ANOVA with independent groups in IBM SPSS, click **Analyze**, then select **General Linear Model** followed by **Univariate**. Notice that we clicked on the same menu items when we conducted the one-way ANOVA in Section 18.10 (see Figure 18.7).

Now, a dialog box will appear. Move the Mean Shock Level Chosen variable (i.e., Shock) to the Dependent Variable box. Then, move both the Alcohol and Expectation variables to the Fixed Factor(s) box. Figure 19.3 shows the dialog box after the variables have been correctly moved. In addition, it provides a screenshot of the data as it would appear in the **Data Editor**. Click OK.

Table 19.3 shows the IBM SPSS output. After the syntax, there is a section of output labeled Between-Subjects Factors. As you can see, it presents the sample size in each level of the factors. The next section of output is labeled Tests of Between-Subjects Effects. In the column labeled Source, we will focus on the rows labeled Alcohol, Expectation, Alcohol *


File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
11 :
var
Univariate
Model...
Contrasts...
Plots...
Post Hoc...
Save...
Options...
OK
Paste
Reset
Cancel
Help

	Shock	Alcohol	Expectation
1	6.00	1.00	1.00
2	8.00	1.00	1.00
3	5.00	1.00	1.00
4	6.00	1.00	1.00
5	7.00	1.00	1.00
6	8.00	1.00	1.00
7	4.00	1.00	1.00
8	9.00	1.00	1.00
9	6.00	1.00	1.00
10	5.00	1.00	1.00
11	5.00	2.00	1.00
12	7.00	2.00	1.00
13	6.00	2.00	1.00
14	5.00	2.00	1.00
15	7.00	2.00	1.00
16	8.00	2.00	1.00
17	4.00	2.00	1.00

Dependent Variable: Mean Shock Leve...

Fixed Factor(s): Alcohol, Expectation

Random Factor(s):

Covariate(s):

WLS Weight:

FIGURE 19.3 IBM SPSS dialog box after selecting variables for a two-factor ANOVA.

TABLE 19.3

```
UNIANOVA Shock BY Alcohol Expectation
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /CRITERIA=ALPHA(0.05)
  /DESIGN=Alcohol Expectation Alcohol*Expectation.
```

Between-Subjects Factors

		Value Label	*N*
Alcohol	1.00	Received alcohol	20
	2.00	Did not receive alcohol	20
Expectation	1.00	Expect alcohol	20
	2.00	Do not expect alcohol	20

Tests of Between-Subjects Effects

Dependent Variable: Mean Shock Level Chosen

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	65.100^a	3	21.700	10.731	.000
Intercept	980.100	1	980.100	484.665	.000
Alcohol	2.500	1	2.500	1.236	.274
Expectation	62.500	1	62.500	30.907	.000
Alcohol * Expectation	.100	1	.100	.049	.825
Error	72.800	36	2.022		
Total	1118.000	40			
Corrected Total	137.900	39			

a. R Squared = .472 (Adjusted R Squared = .428)

Expectation, Error, and Corrected Total. These correspond to Rows, Columns, Rows × Columns, Within cells, and Total, respectively, in Table 19.2. Although all the values across the two tables match one another, notice that the *F* ratio for Expectation in Table 19.3 differs slightly from that shown in Table 19.2. This is due to rounding error.

19.10 Planned Comparisons

planned (*a priori*) comparisons tests used in place of ANOVA when the researcher knows in advance which comparisons are of interest

We do not wish to discourage the comparing and ranking of the original cell means. In fact, the investigator may be primarily concerned with these individual comparisons. As with the one-way design, if theory or logic suggests that particular comparisons are important, then **planned comparisons** are better than omnibus tests of significance such as *F* (see Section 18.16).

Planned comparisons, you recall, are the most powerful tests the investigator can use to discover a difference, if one exists. In a 2 × 2 factorial design, in which there are four treatment groups, we could use the contrasts (1, 1, −1, −1), (1, −1, 1, −1), and (1, −1, −1, 1). These could be described as providing "row," "column," and "interaction" effects, respectively. Be sure to review Section 18.16 carefully.

19.11 Assumptions of the Two-Factor Design and the Problem of Unequal Numbers of Scores

The assumptions for the two-factor ANOVA are the same as for one-way ANOVA: independence of the samples in the cells, random sampling, homogeneity of variance (σ^2 is the same for all cells), and normality of the cell populations (Section 18.11). Violations of the latter two assumptions generally have little effect when there are equal numbers of cases for each level.

In one-way ANOVA, we were able to assign a different number of subjects to the several levels of the treatment variable without serious consequences. However, keeping the number of cases equal is more critical in two-way ANOVA. *The procedures in this chapter assume an equal number of subjects.* If the number of cases per cell is unequal, adjustments have to be made to the data before proceeding. These adjustments, which may reduce power or require some new assumptions that cannot easily be verified, are beyond the scope of this book. Should you need to use these procedures, see Keppel (1991) or Winer, Brown, and Michels (1991).

19.12 Mixed Two-Factor Within-Subjects Design

Air traffic controllers have an enormous responsibility; each must monitor many airplanes as they take off and land at an airport. The job requires that they carefully watch a screen (on which the planes are shown as small blips of light) for long periods of time. When you are flying, your life depends on how well they do their job.

There are many factors that might affect air traffic controllers' performance. One, of course, is lack of sleep. Suppose we have a grant to study the effects of lack of sleep on tasks that require extended vigilance. For our dependent variable, we choose a task in which subjects have to detect targets on a radar screen during a 40-minute session. A subject's score will be number of targets detected. We randomly select four controllers for each of three groups: (1) a group that has been awake for only 2 hours, (2) a group that has been awake for 12 hours, and (3) a group that has had no sleep for 24 hours. Because a controller's job involves numerous sessions with short breaks, we decide to test each subject during four sessions with a 20-minute break between each.

The design we have chosen is a 3 × 4 mixed factorial design. We will test three samples of subjects, one at each level of one factor (Sleep Deprivation), and we will test all the subjects under all four levels of the other factor (Session). This design contains features of both the one-way independent-groups design and the one-way repeated-measures design discussed in the previous chapter. We will test the following null hypotheses:

H_0: $\mu_{A_1} = \mu_{A_2} = \mu_{A_3}$ (Sleep Deprivation has no effect on the subjects' performance)
H_0: $\mu_{B_1} = \mu_{B_2} = \mu_{B_3} = \mu_{B_4}$ (Session has no effect)
H_0: no interaction between Sleep Deprivation and Session in the population

We present some hypothetical data for this study in Table 19.4. Factor *A* is Sleep Deprivation (the independent groups component with three levels) and factor *B* is Sessions (the repeated measures component with four levels). Because the subjects are tested in all four trials, we regard this type of design as a mixed two-factor within-subjects design or an *A* × (*B* × subjects) design. With this design, we can determine whether there is a treatment effect averaged across all the trials (the main effect for *A*), whether Sessions (*B*) has an effect, and whether the effects of Sleep Deprivation are the same for the four sessions (*A* × *B* interaction). One of the major differences between this design and ones we have previously discussed is that there are two error terms needed to form the three *F* ratios. The matrices at the bottom of Table 19.4 give us the information necessary to calculate them.

As usual, we partition the total variation (*SS*) in the data into components attributable to several sources. The first two are identical to the latter two found in the one-way ANOVA with independent groups:

SS_A	Between groups, *A*; derived from the differences among the levels of factor *A*. If *A* has no effect, it is a measure of inherent variation only.
SS_w	Within groups; a measure of inherent variation only.

TABLE 19.4 *Organization Tables for Example of A × (B × Subjects) Design, where Factor A is Sleep Deprivation, Factor B is Session, and the Scores Are the Number of Targets Detected in a Vigilance Task*

ABS

	A_1 (2 hours)					A_2 (12 hours)					A_3 (24 hours)			
	B_1	B_2	B_3	B_4		B_1	B_2	B_3	B_4		B_1	B_2	B_3	B_4
S_1	40	42	45	45	S_5	38	38	34	33	S_9	34	34	30	28
S_2	36	40	38	39	S_6	40	41	37	36	S_{10}	36	34	31	26
S_3	46	46	44	42	S_7	44	42	42	39	S_{11}	32	34	30	28
S_4	39	40	36	40	S_8	41	43	40	37	S_{12}	34	30	28	25

A × S

	A_1		A_2		A_3
S_1	172	S_5	143	S_9	126
S_2	153	S_6	154	S_{10}	127
S_3	178	S_7	167	S_{11}	124
S_4	155	S_8	161	S_{12}	117
Sum	658	Sum	625	Sum	494

A × B

	B_1	B_2	B_3	B_4	Sum
A_1	161	168	163	166	658
A_2	163	164	153	145	625
A_3	136	132	119	107	494
Sum	460	464	435	418	1777

The next three are similar to those found in the repeated-measures design:

SS_B	Between trials, *B*; derived from the differences among the trials.
$SS_{A\times B}$	$A \times B$ interaction; the joint effect of *A* and *B*.
$SS_{B\times \mathrm{w}}$	Residual; the same as in the one-way design except that there are different subjects in each level of factor *A*.

The computational formulas for the sums of squares, with worked examples from Table 19.4, are as follows (a = number of levels of *A*, b = number of levels of *B*):

TOTAL SUM OF SQUARES:

$$SS_{\text{total}} = \sum^{\substack{\text{all}\\ \text{scores}}} X^2 - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X\right)^2}{n_{\text{total}}} \tag{19.10}$$

$$\begin{aligned} SS_{\text{total}} &= [40^2 + 42^2 + \ldots 28^2 + 25^2] - \frac{1777^2}{48} \\ &= 67{,}201 - 65{,}786.02 \\ &= 1414.98 \end{aligned}$$

BETWEEN-GROUPS
SUM OF SQUARES:

$$SS_A = \frac{\sum A^2}{(n_A)(b)} - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X\right)^2}{n_{\text{total}}} \tag{19.11}$$

$$\begin{aligned} SS_A &= \frac{\left(\sum X_{A_1}\right)^2 + \left(\sum X_{A_2}\right)^2 + \left(\sum X_{A_3}\right)^2}{(n_A)(b)} - \frac{\left(\sum^{\substack{\text{all}\\ \text{scores}}} X\right)^2}{n_{\text{total}}} \\ &= \frac{658^2 + 625^2 + 494^2}{(4)(4)} - \frac{1777^2}{48} \\ &= 66{,}726.56 - 65{,}786.02 \\ &= 940.54 \end{aligned}$$

WITHIN-GROUPS
SUM OF SQUARES:

$$SS_{\mathrm{w}} = \frac{\sum (A \times S)^2}{b} - \frac{\sum A^2}{(n_A)(b)} \tag{19.12}$$

$$SS_w = \frac{\left(\sum X_{S_1}\right)^2 + \left(\sum X_{S_2}\right)^2 + \ldots + \left(\sum X_{S_{12}}\right)^2}{b} - \frac{\left(\sum X_{A_1}\right)^2 + \left(\sum X_{A_2}\right)^2 + \left(\sum X_{A_3}\right)^2}{(n_B)(b)}$$

$$= \frac{172^2 + 153^2 + \cdots + 117^2}{4} - \frac{658^2 + 625^2 + 494^2}{16}$$

$$= 66{,}936.75 - 66{,}726.56$$

$$= 210.19$$

BETWEEN-TRIALS
SUM OF SQUARES:

$$SS_B = \frac{\sum B^2}{(n_A)(a)} - \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}} \qquad (19.13)$$

$$SS_B = \frac{\left(\sum X_{B_1}\right)^2 + \left(\sum X_{B_2}\right)^2 + \left(\sum X_{B_3}\right)^2}{(n_A)(a)} - \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}}$$

$$= \frac{460^2 + 464^2 + 435^2 + 418^2}{(4)(3)} - \frac{1777^2}{48}$$

$$= 65{,}903.75 - 65{,}786.02$$

$$= 117.73$$

INTERACTION
SUM OF SQUARES:

$$SS_{A\times B} = \frac{\sum (A \times B)^2}{n_A} - \frac{\sum A^2}{(n_A)(b)} - \frac{\sum B^2}{(n_A)(a)} + \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}} \qquad (19.14)$$

$$SS_{A\times B} = \frac{\left(\sum X_{A_1B_1}\right)^2 + \left(\sum X_{A_1B_2}\right)^2 + \ldots + \left(\sum X_{A_3B_4}\right)^2}{n_a} - \frac{\sum A^2}{(n_A)(b)} - \frac{\sum B^2}{(n_A)(a)} + \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}}$$

$$= \frac{161^2 + 168^2 + \cdots + 107^2}{4} - 66{,}726.56 - 65{,}903.75 + 65{,}786.02$$

$$= 66{,}924.75 - 66{,}726.56 - 65{,}903.75 + 65{,}786.02$$

$$= 80.46$$

RESIDUAL SUM OF SQUARES:

$$SS_{B\times \mathrm{w}} = \sum^{\substack{\text{all}\\ \text{scores}}} X^2 - \frac{\sum (A \times B)^2}{n_A} - \frac{\sum (A \times S)^2}{b} + \frac{\sum A^2}{(n_A)(b)} \tag{19.15}$$

$$SS_{B\times \mathrm{w}} = 67{,}201 - 66{,}924.75 - 66{,}936.75 + 66{,}726.56$$

$$= 66.06$$

As usual, $SS_{\text{total}} = SS_A + SS_{\mathrm{w}} + SS_B + SS_{A\times B} + SS_{B\times \mathrm{w}}$. We show the general form of the summary table at the top of Table 19.5 and our worked example at the bottom. You can find the formulas for determining the degrees of freedom in the table. Notice that we use s^2_{w} in the denominator to test for the main effect of A, but we use $s^2_{B\times \mathrm{w}}$ in the denominator for the other two F ratios.

TABLE 19.5 *General Form and Worked Example of the Summary Table for an $A \times (B \times \text{Subjects})$ Design*

SOURCE	SS	df	s^2	F
A	Formula 19.11	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{s^2_A}{s^2_{\mathrm{w}}}$
Within	Formula 19.12	$(a)(n_A - 1)$	$\frac{SS_{\mathrm{w}}}{df_{\mathrm{w}}}$	
B	Formula 19.13	$b - 1$	$\frac{SS_B}{df_B}$	$\frac{s^2_B}{s^2_{B\times \mathrm{w}}}$
$A \times B$	Formula 19.14	$(a - 1)(b - 1)$	$\frac{SS_{A\times B}}{df_{A\times B}}$	$\frac{s^2_{A\times B}}{s^2_{B\times \mathrm{w}}}$
$B \times$ within	Formula 19.15	$(a)(b - 1)(n_A - 1)$	$\frac{SS_{B\times \mathrm{w}}}{df_{B\times \mathrm{w}}}$	
Total	Formula 19.10	$n_{\text{total}} - 1$		

SOURCE	SS	df	s^2	F
A	940.54	2	470.27	20.14
Within	210.19	9	23.35	
B	117.73	3	39.24	16.02
$A \times B$	80.46	6	13.41	5.47
$B \times$ within	66.06	27	2.45	
Total	1414.98	47		

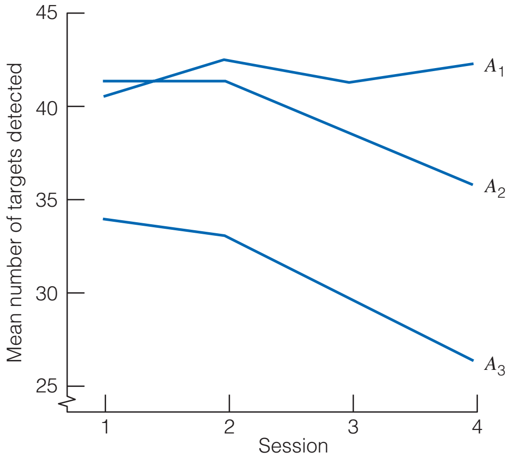


FIGURE 19.4 Results of the $A \times (B \times \text{subjects})$ design, where A_1, A_2, and A_3 indicate levels of factor A (sleep deprivation).

In our example, all three effects were statistically significant. We show a graph of the results in Figure 19.4. We can conclude that sleep deprivation affected performance, and that performance differed over trials. Because there are more than two levels for factors, we must use *post hoc* tests to determine what levels were significantly different. The significant interaction, however, indicates that the effect of time without sleep on performance depends on the trial as well.

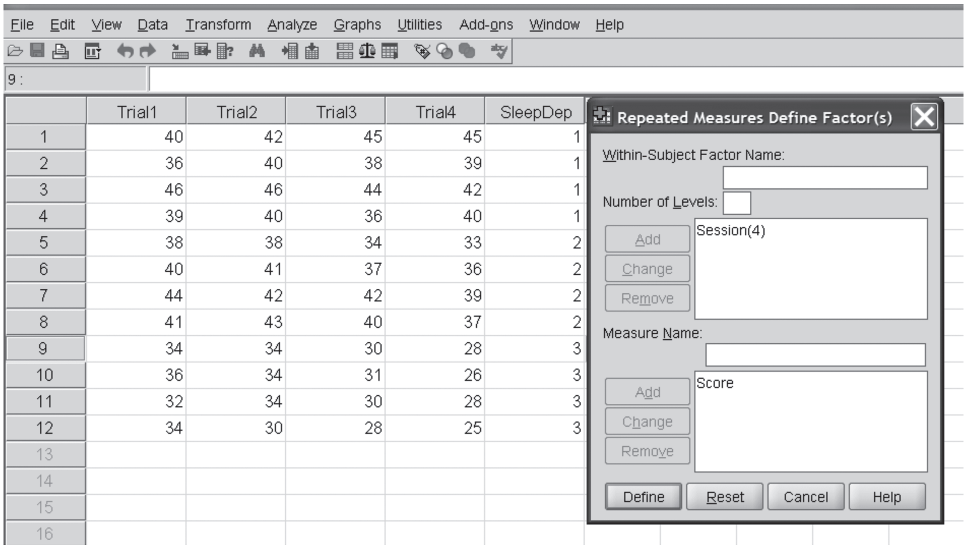
We have given a very brief treatment of the mixed two-factor within-subjects design. It is not a simple analysis, and some advanced textbooks devote an entire chapter to the analysis of the interaction alone (Keppel, 1991). Our intention is only to give you a feel for the more complex designs possible with ANOVA. You no doubt can appreciate the amount of work that would be required if you had to make all calculations by hand. Fortunately, we can use IBM SPSS to assist us with using complex designs.

19.13 Calculation of the Mixed Two-Factor Within-Subjects Design with IBM SPSS

In this section, you will learn how to use IBM SPSS to analyze a mixed two-factor within-subjects design. You will use the data from Table 19.4.

In the **Data Editor**, all the scores for the dependent variable (i.e., number of targets detected) should be entered in four different columns—one for each of the four trials of the repeated measures. This should be done for each of the 12 subjects. In the fifth column, we need to identify from what condition the scores came. We will enter 1s for scores of subjects who were awake for 2 hours, 2s for scores of subjects who were awake for 12 hours, and 3s for scores of subjects who were awake for 24 hours. In the **Variable View**, for the first four variables, we will use Trial1, Trial2, Trial3, and Trial4, respectively, as the variable names. For the fifth column, we will use SleepDep for the variable name and Sleep Deprivation as the label. For all variables, we will also change the number of decimals displayed from 2 to 0 (see Section 7.8). Similar to Section 14.7 and Section 18.10, be sure to use the Labels column to enter value labels for SleepDep.

To conduct a mixed two-factor within subjects design in IBM SPSS, click **Analyze**, then select **General Linear Model** followed by **Repeated Measures**. Notice that we clicked



	Trial1	Trial2	Trial3	Trial4	SleepDep
1	40	42	45	45	1
2	36	40	38	39	1
3	46	46	44	42	1
4	39	40	36	40	1
5	38	38	34	33	2
6	40	41	37	36	2
7	44	42	42	39	2
8	41	43	40	37	2
9	34	34	30	28	3
10	36	34	31	26	3
11	32	34	30	28	3
12	34	30	28	25	3
13					
14					
15					
16					

Within-Subject Factor Name:

Number of Levels:

Session(4)

Measure Name:

Score

FIGURE 19.5 IBM SPSS dialog box when specifying repeated measures in ANOVA.

on the same menu items when we conducted the one-way repeated measures ANOVA in Section 18.19 (see Figure 18.10).

Now, a dialog box will appear. In the box labeled Within-Subject Factor Name, type "Session," and for Number of Levels, type "4." Click Add. Then, in the box labeled Measure Name, type "Score." Figure 19.5 shows the dialog box after the labels and levels have been correctly entered, including the data as it would appear in the **Data Editor**. Click Define. Now, a new dialog box will appear. You will simply move the four variables representing the repeated measures (i.e., Trial1, Trial2, Trial3, and Trial4) to the box labeled Within-Subjects Variables. Then, move the Sleep Deprivation variable to the box labeled Between-Subjects Factor(s). Click OK.

Because IBM SPSS produces numerous tables (much of which is beyond the scope of this book), we show only an excerpt of the IBM SPSS output in Table 19.6. After the syntax, we present two sections of output labeled Tests of Within-Subjects Effects and Tests of Between-Subjects Effects, respectively. In the Tests of Within-Subjects Effects, for the row labeled Session, Sphericity Assumed, this corresponds to the row labeled B at the bottom of Table 19.5. For the row in the IBM SPSS output labeled Session * SleepDep, Sphericity Assumed, this corresponds to the row labeled $A \times B$ in Table 19.5. For the row in the IBM SPSS output labeled Error(Session), Sphericity Assumed, this corresponds to the row labeled $B \times$ within at the bottom of Table 19.5. See Section 18.18 and Section 18.19 for a discussion of the sphericity assumption. In the IBM SPSS output labeled Tests of Between-Subjects Effects, the rows labeled SleepDep and Error, respectively, correspond to A and Within in Table 19.5. You will notice that the three F statistics in Table 19.5 match those values in the IBM SPSS output within rounding error.

19.14 Summary

In this chapter, we extended ANOVA to situations involving two treatment variables. The two variables are called factors, and the various conditions of each are referred to as levels. One advantage of the two-way design is that we can use the same subjects to examine the effects of both variables simultaneously. A second advantage is that our conclusions can be broader—we can make conclusions about each variable for all the conditions (levels) of the other.

TABLE 19.6 *IBM SPSS Output From a Mixed Two-Factor Within-Subjects Design*

```
GLM Trial1 Trial2 Trial3 Trial4 BY SleepDep
  /WSFACTOR=Session 4 Polynomial
  /MEASURE=Score
  /METHOD=SSTYPE(3)
  /CRITERIA=ALPHA(.05)
  /WSDESIGN=Session
  /DESIGN=SleepDep.
```

Tests of Within-Subjects Effects

Measure: Score

Source		Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Session	Sphericity Assumed	117.729	3	39.243	16.039	.000
	Greenhouse-Geisser	117.729	2.370	49.679	16.039	.000
	Huynh-Feldt	117.729	3.000	39.243	16.039	.000
	Lower-bound	117.729	1.000	117.729	16.039	.003
Session *SleepDep	Sphericity Assumed	80.458	6	13.410	5.481	.001
	Greenhouse-Geisser	80.458	4.740	16.976	5.481	.002
	Huynh-Feldt	80.458	6.000	13.410	5.481	.001
	Lower-bound	80.458	2.000	40.229	5.481	.028
Error (Session)	Sphericity Assumed	66.063	27	2.447		
	Greenhouse-Geisser	66.063	21.328	3.097		
	Huynh-Feldt	66.063	27.000	2.447		
	Lower-bound	66.063	9.000	7.340		

Tests of Between-Subjects Effects

Measure: Score
Transformed Variable: Average

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Intercept	65786.021	1	65786.021	2816.886	.000
SleepDep	940.542	2	470.271	20.136	.000
Error	210.188	9	23.354		

In the two-factor ANOVA with independent groups, we test three null hypotheses about the means of various populations: (1) the hypothesis of no difference among the levels of the row factor, (2) the hypothesis of no difference among the levels of the column factor, and (3) the hypothesis of no interaction between the row and column factors. For the first two, a significant F ratio is referred to as a main effect—a finding of differences among the levels of that factor *averaged* across the levels of the other factor. Thus, the F test for a main effect is analogous to a one-way ANOVA.

Interaction refers to the joint effect of the two independent variables (factors) on the dependent variable. A significant interaction effect indicates that the effect of one factor differs from level to level of the second factor, and vice versa. If interaction is present, the main effect of a factor can be misleading, for it will not apply equally at all levels of the other factor. The ability to examine interaction is the third important advantage of the two-way design over the one-way design.

In this chapter, we also introduced the mixed two-factor within-subjects design, which combines features of the one-way independent-groups design and the one-way repeated-measures design. With this design, different subjects are selected for the different levels of one factor, but all the subjects are tested under all the levels of the other factor.

Key Terms, Concepts, and Symbols

factor (p. 336)
two-factor analysis of variance design (p. 336)
levels (p. 336)
cell (p. 336)
main effect (p. 337)
main effect null hypothesis (p. 337)
interaction (p. 338)
ordinal interaction (p. 340)
disordinal interaction (p. 340)
null hypothesis of no interaction (p. 340)
simple effect (p. 341)
SS_{total} (p. 342)
SS_{wc} (p. 342)
SS_R (p. 343)
SS_C (p. 343)
$SS_{R \times C}$ (p. 344)
df_{total}, df_{wc}, df_R, df_C, $df_{R \times C}$ (p. 346)
σ^2 (p. 346)
s^2_{wc}, s^2_R, s^2_C, $s^2_{R \times C}$ (pp. 346, 347)
Tukey's HSD test (p. 348)
η^2 (p. 349)
partial omega-squared (p. 349)
$\hat{\omega}^2$ (p. 349)
planned comparisons (p. 352)
mixed two-factor within-subjects design (p. 352)
A × (B × subjects) design (p. 353)
s^2_A, s^2_{w}, s^2_B, $s^2_{A \times B}$, $s^2_{B \times \text{w}}$ (p. 356)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Suppose an educational psychologist is carrying out an experiment in which grade school children hear one of four stories: (1) a stereotypical (and sexist) story in which a boy is the active hero while a girl just passively watches him, (2) a reverse-stereotype story in which a girl is active while a boy just watches, (3) a story in which both a boy and a girl are active, or (4) a story in which both a boy and a girl are passive. Of the children who hear a given story, 15 are boys and 15 girls. After hearing the story (whichever of the four kinds it is), each subject is given an opportunity to work at the mildly frustrating task of using long-handled spoons to rearrange a terrarium in a narrow-necked jar. The psychologist observes how long the subject persists at the task. This experiment has which kind of design? (a) A one-way design in which type of story is the treatment variable. (b) A one-way design in which sex of subject is the treatment variable. (c) A two-way design in which type of story and sex of subject are the treatment variables. (d) A two-way design in which type of story and persistence at the task are the treatment variables. Explain.

2. (a) In the experiment described in Problem 1, how many levels does the factor Type of Story have? (b) How many cells are there? (c) Explain what a finding of main effect for Type of Story means here.

3. Which of the following outcomes for the experiment described in Problem 1 would indicate interaction between Type of Story and Sex of Subject for the samples on hand? (a) The mean persistence score for all boys equals the mean persistence score for all girls. (b) The mean persistence score for all children who hear a given story is the same for all four stories. (c) The mean persistence score for all boys is 18.3 sec longer than the mean persistence score for all girls. (d) The mean persistence score for boys is greater than that for girls when the story the child just heard featured an active male character, but it is lower than the mean persistence score for girls when the story featured an active female character. Explain.

4. "A good way to begin the computation of a two-way ANOVA is to find for each cell the number of raw scores, the sum of them, and the sum of their squares." Would you agree with this statement? Explain.

5. An experimenter wishes to evaluate the effectiveness of four teaching methods among three kinds of students. The study is conducted as a two-way ANOVA, and each combination of method and type of student is represented by five subjects. (a) What *df* characterizes each of the *SS*s that will be calculated? (b) If the tests are conducted at the 5% level of significance, what is the critical value of *F* for testing difference among methods? For testing difference among types of students? For testing interaction?

6. Relative to Data 19A, suppose that 5 subjects had been assigned to each cell. (a) State the values of $df_{R \times C}$, df_{wc}, and df_{total}. (b) Make the possible *F* tests, using $\alpha = .05$. (c) What can we say about the meaning of the outcome of the test for row effect?

DATA 19A

SOURCE	***df***	s^2	***F***
Rows	3	36.0	
Columns	4	18.0	
Rows × Columns		30.0	
Within cells		12.0	
Total			

7. Examine the graphs of the cell means from three 2 × 3 analyses of variance and one 3 × 3 ANOVA shown in Figure 19.6. Suppose that in each case all three *F* tests (row, column, and interaction) proved to be significant. (a) Which interactions are ordinal and which are disordinal? (b) In which case(s) would conclusions based on a main effect for row factor alone be misleading? Explain.

8. An experiment was designed to test the relative effectiveness of a computer-taught method of instruction compared with a conventional lecture method. Two samples of three subjects each were selected from among students of good aptitude and assigned to the two methods of instruction, as shown in Data 19B. Two samples of the same

DATA 19B

	Computer-taught	*Lecture method*
Bright students	5, 7, 9	10, 11, 15
Ordinary students	5, 10, 9	4, 8, 6

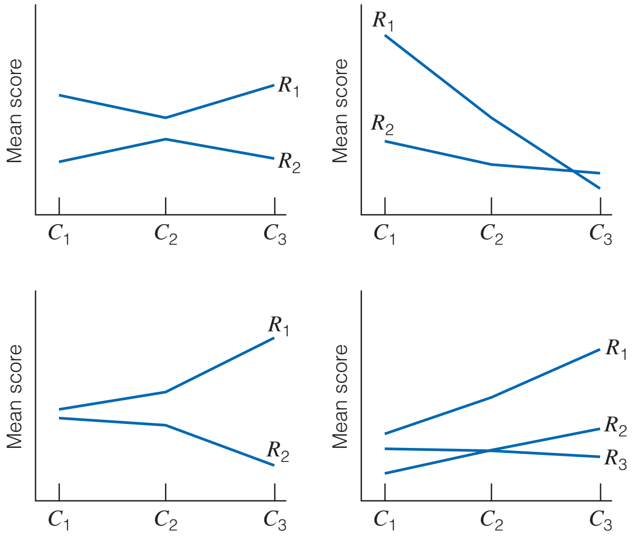


FIGURE 19.6

size were selected from among students of ordinary ability and assigned to the same methods of instruction (also shown in Data 19B). In the test trial, the 12 students earned the scores shown in Data 19B. (a) Find df_{total}, df_R, df_C, df_{wc}, and $df_{R \times C}$. (b) Find SS_{total}, SS_R, SS_C, SS_{wc}, and $SS_{R \times C}$. (c) Calculate the several variance estimates appropriate to a two-way ANOVA, and present the results thus far developed in an analysis of variance table. (d) Calculate the Fs required to test the main effects and the interaction effect. Draw the statistical conclusions appropriate when $\alpha = .05$ and $\alpha = .01$. (e) Calculate the effect size, $\hat{\omega}^2$. (f) Construct a graph of the cell means in the manner of Figure 19.1. (g) In light of the statistical outcome (d) and the graph (e), what appears to be the essential outcome of the study?

9. In a learning experiment, we want to know whether there is a difference in performance among three kinds of motivation and also whether there is a difference between two conditions of practice. Data 19C shows the 12 scores that might be obtained in such a study, if two subjects were assigned at random to each of the six possible combinations of motivation and condition of practice. (a) Complete the two-way ANOVA with summary table. (b) Calculate the effect sizes, $\hat{\omega}^2$. (c) Construct a graph of the cell means.

Data 19C

		Kind of motivation		
		C_1	C_2	C_3
Condition of practice	R_1	4, 4	2, 3	8, 6
	R_2	6, 8	5, 4	10, 13

10. For Data 19B (Problem 8), calculate the corrected cell means necessary to interpret the interaction effect.

11. For the data in Problem 9, use IBM SPSS to calculate a two-way ANOVA.

12. We wish to examine the effects of food deprivation (factor A) on motivated behavior. We randomly assign 15 rats to one of three groups: (1) the rats are deprived of food for 12 hours, (2) they are deprived of food for 24 hours, and (3) they are deprived of food for 48 hours. As our dependent variable, we record the time (in seconds) it takes each rat to traverse a 10-foot runway for food reinforcement on three consecutive trials (factor B). None of the animals have any previous experience in the runway. Suppose that the results are as given in Data 19D. (a) complete the two-way ANOVA and present the results in a summary table. (b) Construct a graph of the results. (c) State your conclusions.

Data 19D

	A_1 (12 HOURS)				A_2 (24 HOURS)				A_3 (48 HOURS)		
	B_1	B_2	B_3		B_1	B_2	B_3		B_1	B_2	B_3
S_1	60	40	25	S_6	40	30	15	S_{11}	50	20	5
S_2	50	30	20	S_7	60	25	10	S_{12}	60	15	7
S_3	40	35	20	S_8	55	20	15	S_{13}	40	20	10
S_4	60	30	30	S_9	45	25	8	S_{14}	50	15	8
S_5	70	40	15	S_{10}	55	35	9	S_{15}	55	10	10

CHAPTER 20

Chi-Square and Inference about Frequencies

When you have finished studying this chapter, you should be able to:

- Understand that the chi-square test is used to test hypotheses about the number of cases falling into the categories of a frequency distribution;
- Understand that χ^2 provides a measure of the difference between observed frequencies and the frequencies that would be expected if the null hypothesis were true;
- Explain why the chi-square test is best viewed as a test about proportions;
- Compute χ^2 for one-variable goodness-of-fit problems;
- Compute χ^2 to test for independence between two variables; and
- Compute effect size for the chi-square test.

In previous chapters, we have been concerned with numerical scores and testing hypotheses about the mean or the correlation coefficient. In this chapter, you will learn to make inferences about frequencies—the number of cases falling into the categories of a frequency distribution. For example, among four brands of soft drinks, is there a difference in the proportion of consumers who prefer the taste of each? Is there a difference among registered voters in their preference for three candidates running for local office? To answer questions like these, a researcher compares the observed (sample) frequencies for the several categories of the distribution with those frequencies expected according to his or her hypothesis. The difference between observed and expected frequencies is expressed in terms of a statistic named *chi-square* (χ^2), introduced by Karl Pearson in 1900.

20.1 The Chi-Square Test for Goodness of Fit

categorical data
data comprising qualitative categories

P
symbol for proportion

The chi-square (pronounced "ki") test was developed for **categorical data**; that is, for data comprising qualitative categories, such as eye color, gender, or political affiliation. Although the chi-square test is conducted in terms of frequencies, it is best viewed conceptually as a test about proportions. For example, let us return to the question of whether four brands of soft drinks are equally preferred by consumers. Here, we explore the possibility that in the population of consumers, the proportion of individuals preferring each brand is 1/4:

$$H_0\colon P_A = P_B = P_C = P_D = .25$$

H_0 for chi-square test for goodness of fit
states the *proportion* of cases in each category of a population of qualitative data

The alternative hypothesis is simply that the null hypothesis is untrue in some (any) way. It could be that a preference exists for the first of the brands and the remainder are equally less attractive, or that the first two are preferred more than the second two, and so on.

goodness-of-fit problem
asks whether the relative frequencies observed in the categories of a sample frequency distribution are in agreement with the relative frequencies expected to be true in the population (as stated in H_0)

Our hypothetical soft drink problem is an example of a class of problems commonly referred to as **goodness-of-fit problems**. The use of chi-square tells us whether the relative frequencies observed in the two or more categories of our sample frequency distribution agree with the set of frequencies hypothesized (expected) to be true of the population distribution. For this type of problem, chi-square is a *one-variable* test with k number of categories.

k
symbol for number of categories

To test the null hypothesis in our soft drink study, we might allow subjects to taste each of the four brands and then declare their preference. We would, of course, control for possible extraneous influences, such as knowledge of brand name or order of presentation. Suppose we randomly select 100 subjects and that our **observed frequencies (f_o)** of preference are as shown in Table 20.1.

An **expected frequency (f_e)** is the mean of the observed frequencies that would occur on infinite repetitions of an experiment when sampling is random and the null hypothesis is true. (In our example, the null hypothesis is true when there is no preference among consumers for the four soft drinks.) We calculate the expected frequency for each category by multiplying the proportion hypothesized to characterize that category in the population by the sample size. According to the null hypothesis, in our example, the expected proportionate preference for each soft drink is 1/4, and the expected frequency of preference for each is therefore (1/4)(100) = 25.

TABLE 20.1 *Observed and Expected Frequency of Preference for Four Brands of Soft Drinks (100 Subjects)*

	Brand A	*Brand B*	*Brand C*	*Brand D*	
Observed frequency	$f_o = 15$	$f_o = 36$	$f_o = 28$	$f_o = 21$	$\sum f_o = 100$
Expected frequency	$f_e = 25$	$f_e = 25$	$f_e = 25$	$f_e = 25$	$\sum f_e = 100$

observed frequencies (f_o) numbers of observations within a particular category of a sample of qualitative data

expected frequency (f_e) numbers of observations within a particular category that is predicted under the condition that the null hypothesis is true

OPINION POLLS INC.
D.Fradon

"How would you like me to answer that question? As a member of my ethnic group, educational class, income group, or religious category?"

© *The New Yorker Collection,* 1969. Dana Fradon, from cartoonbank.com. All Rights Reserved.

In any one experiment, we anticipate that the observed frequency of choices will vary from the expected frequencies because of chance (random sampling variation). In our example, even if the null hypothesis is correct and there is no preference among consumers in the population, (and thus the expected frequencies for the four soft drinks are 25, 25, 25, and 25), would you be surprised if in a sample of 100 persons randomly drawn from the population, the observed frequencies were A = 27, B = 23, C = 24, and D = 26? Or A = 24, B = 28, C = 22, and D = 26? Probably not, because you expect some variation due to chance factors. But how much variation is reasonable to expect? Some measure is needed to test whether the observed differences did or did not have a low probability of occurring by chance if the null hypothesis is true.

20.2 Chi-Square (χ^2) as a Measure of the Difference between Observed and Expected Frequencies

chi-square statistic provides a measure of the difference between observed and expected frequencies

The **chi-square statistic, χ^2**, provides a measure of the difference between observed and expected frequencies. Its basic formula is:

FORMULA FOR CHI-SQUARE

$$\chi^2 = \sum\left(\frac{(f_o - f_e)^2}{f_e}\right) \tag{20.1}$$

where: f_o = the observed frequency. Summation is over the number of categories characterizing a given problem.
f_e = the expected frequency.

χ^2
symbol for chi-square

$\sum\left(\frac{(f_o - f_e)^2}{f_e}\right)$
formula for the chi-square statistic

For the data of Table 20.1 there are four categories (brands), and we calculate chi-square as follows:

$$\chi^2 = \frac{(15-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(28-25)^2}{25} + \frac{(21-25)^2}{25}$$

$$= \frac{(-10)^2}{25} + \frac{11^2}{25} + \frac{3^2}{25} + \frac{(-4)^2}{25}$$

$$= 4.0 + 4.84 + .36 + .64 = 9.84$$

Examine the formula and the calculation; they reveal several points of interest about χ^2:

1. χ^2 cannot be negative because all differences are squared; both positive and negative differences make a positive contribution to the value of χ^2.

2. χ^2 will be zero only in the unusual event that each observed frequency exactly equals the corresponding expected frequency.

3. Other things being equal, the larger the difference between the values of f_o and their corresponding values of f_e, the larger χ^2.

4. It is not the size of the difference alone that accounts for a contribution to the value of χ^2; rather, it is the size of the difference *relative to the magnitude of the expected frequency.*

This is intuitively reasonable. For example, if we toss a number of coins and inquire as to the number of "heads," f_e is 6 when we toss 12 coins and 500 when we toss 1,000 coins. If we obtain 11 heads in 12 tosses, the difference is 5—most unusual. However, if we obtain 505 heads in 1,000 tosses, the difference is also 5—but hardly a rarity.

5. The value of χ^2 depends on the *number* of differences involved in its calculation.

$k - 1$
degrees of freedom in a goodness-of-fit problem

For example, if we used three brands of soft drinks instead of four, there would be one less difference to contribute to the total of χ^2. We take this factor into account by considering the number of **degrees of freedom (*df*)** associated with the particular χ^2. For this type of problem, the degrees of freedom will be $k - 1$ (because only the values of $k - 1$ categories are free to vary for a fixed value of χ^2). In our soft drink problem, then, $df = 4 - 1 = 3$.

20.3 The Logic of the Chi-Square Test

In the chi-square test, the null hypothesis is that in the *population*, the proportional frequency of cases in each category equals a specified value. The proportions hypothesized are derived from a research question of interest to the researcher.

When the null hypothesis is true *and* certain conditions hold (see Section 20.7), the sampling distribution formed by the values of χ^2 calculated from repeated random samples closely follows a known theoretical distribution. Actually, there is a *family* of sampling distributions of χ^2, each member corresponding to a given number of degrees of freedom. Figure 20.1 shows the sampling distributions for several different values of *df.*

Suppose that the null hypothesis is true. Then the observed frequencies will differ from their corresponding expected frequencies only because of chance (random sampling variation). The calculated value of χ^2 will be small when the difference between the values of f_o and the values of f_e is small and large when it is not. To test the null hypothesis, we must learn what calculated values of χ^2 would occur under random sampling when the null hypothesis is true. Then, we will compare the χ^2 calculated from our particular sample with

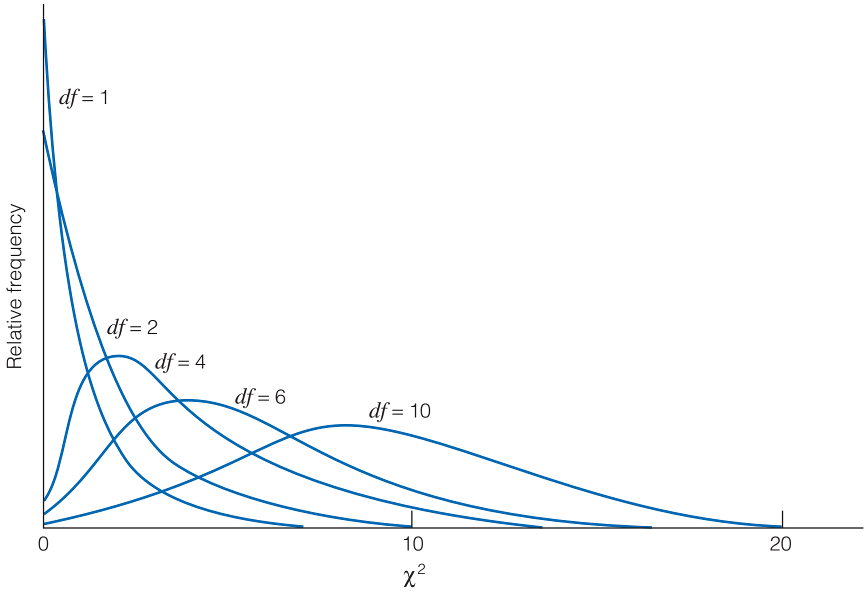


FIGURE 20.1 Approximate forms of the χ^2 distribution for various degrees of freedom. (*Source:* E. F. Lindquist, *Design and Analysis of Experiments in Psychology and Education.* Copyright © 1953 by Houghton Mifflin Company. Used with permission.)

this distribution of values. If it is so large that such a value would rarely occur when the null hypothesis is true, we will reject the hypothesis.

The region of rejection always appears in the upper tail of the sampling distribution of χ^2. This is because we square differences between observed and expected frequencies; thus, larger differences always result in a numerically higher value of χ^2_{calc}.

To conduct the test, we must determine the critical value of χ^2. At the 5% significance level, this is the value of χ^2 so large that it will be equaled or exceeded only 5% of the time in random sampling when the null hypothesis is true. We use Table I in Appendix D to determine the critical value for the desired level of significance and degrees of freedom. As usual, we reject the null hypothesis if $\chi^2_{\text{calc}} \geq \chi^2_{\text{crit}}$.

In our soft drink problem, $df = 3$. Entering Table I with this value, we find that the critical value is 7.81 when $\alpha = .05$. We illustrate this test situation in Figure 20.2. Our calculated value of χ^2 is 9.84, which exceeds the critical value. The $(f_o - f_e)$ differences are therefore so great that they have a low probability of occurring by chance if the null hypothesis is true, so we reject the null hypothesis.

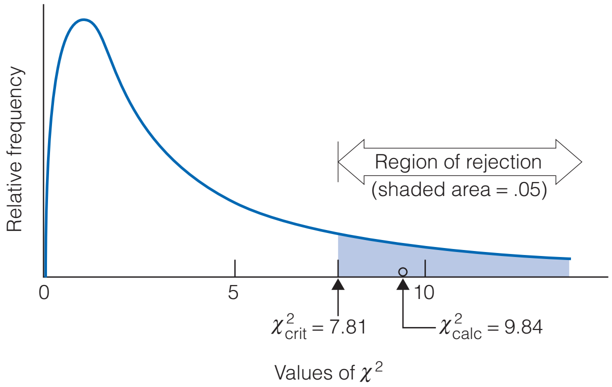


FIGURE 20.2 Location of the calculated value of χ^2 relative to the expected distribution of χ^2 when $df = 3$.

What about the other tail of the chi-square distribution? A very low obtained value of χ^2 means that, *if the necessary assumptions have been met and the calculations are correct,* the observed frequencies are closer to the expected frequencies than chance would ordinarily allow. A very small value of χ^2 has no significance other than to alert us to check on the adequacy of our procedures.

20.4 Interpretation of the Outcome of a Chi-Square Test

If we find that χ^2 is significant, we have only opened the door to explanatory analysis. In our soft drink experiment we obtained a sample χ^2 larger than the critical value of 7.81, so we rejected H_0 and concluded that some soft drinks are preferred over others. Which ones? We are tempted to say that people prefer brand B. Strictly speaking, however, we cannot tell from our χ^2 value alone because H_A, the alternative hypothesis, is simply that H_0 is untrue in some (any) way, and there are many ways in which that could occur. It may be that people equally prefer brands B and C, and that both are preferred over brands A and D. Note that the distinction between a directional test and a nondirectional test is not pertinent here; a directional test is possible only in situations where there are only two ways (directions) in which H_0 can be wrong.

When $df = 1$ (i.e., when there are only two categories), we can view the chi-square test as a *test about a single proportion.* In this case, it is possible to perform a one-tailed test if the logic of the study calls for it. When $df = 1$ (only), $\chi^2 = z^2$. We may therefore calculate $z = \sqrt{\chi^2_{\text{calc}}}$ and compare our result with the critical one-tailed value of the normally distributed z ($z_{\text{crit}} = 1.645$ for $\alpha = .05$ and $z_{\text{crit}} = 2.33$ for $\alpha = .01$). We reject the null hypothesis only for differences in the direction specified in the alternative hypothesis.

20.5 Different Hypothesized Proportions in the Test for Goodness of Fit

In any goodness-of-fit problem, the hypothesized proportion in each category is dictated by the research question of interest. In our soft drink problem, the hypothesized proportions reflected the possibility that the four brands are equally attractive. However, the hypothesized (expected) proportions do not have to be equal, as we will now show.

Chi-square is frequently used in the field of population genetics. Suppose, for example, that a geneticist believes that plants bred in a particular way will show one of four characteristics according to the ratio 9:3:3:1. In an experiment designed to test this hypothesis, the geneticist converts these ratios into expected proportions: $P_1 = 9/16$, $P_2 = 3/16$, $P_3 = 3/16$, and $P_4 = 1/16$. If the geneticist breeds plants in the specified manner and obtains a sample of 256 seedlings, the expected frequencies are 144, 48, 48, and 16, respectively (obtained by multiplying each of the expected proportions by 256). Suppose the geneticist obtained the hypothetical results shown in Table 20.2. By Formula 20.1, the calculated value of χ^2 is:

$$\chi^2 = \frac{(130 - 144)^2}{144} + \frac{(55 - 48)^2}{48} + \frac{(50 - 48)^2}{48} + \frac{(21 - 16)^2}{16} = 4.03$$

Is the difference between expected and observed frequencies so great that it has a low probability of occurring by chance? Here, $df = k - 1 = 3$. We turn to Table I in Appendix D and find that when $df = 3$, a χ^2 of 7.81 or greater would occur by chance 5% of the time when

TABLE 20.2 *Observed and Expected Frequency of Occurrence of Four Inherited Plant Characteristics*

	Type 1	*Type 2*	*Type 3*	*Type 4*	
Observed frequency	$f_o = 130$	$f_o = 55$	$f_o = 50$	$f_o = 21$	$\sum f_o = 256$
Expected frequency	$f_e = 144$	$f_e = 48$	$f_e = 48$	$f_e = 16$	$\sum f_e = 256$

the null hypothesis is true. The geneticist's χ^2 is not so rare an event (when $\alpha = .05$) to give her a reason to reject the null hypothesis. The sample outcome is one of many that could have easily occurred by chance when the null hypothesis is true.

20.6 Effect Size for Goodness-of-Fit Problems

$\hat{\omega}$
omega; a measure of effect size for the chi-square goodness-of-fit test

Jacob Cohen (1988) proposes that the appropriate measure of effect size for the chi-square goodness-of-fit test is $\hat{\omega}$ (omega):

$$\sqrt{\sum\left[\frac{(p_o - p_e)^2}{p_e}\right]}$$
formula for $\hat{\omega}$

FORMULA FOR EFFECT SIZE, $\hat{\omega}$

$$\hat{\omega} = \sqrt{\sum\left[\frac{(p_o - p_e)^2}{p_e}\right]} \qquad (20.2)$$

where: p_o = observed proportion
p_e = expected proportion.

Omega is a member of the *d* family of effect sizes (see Section 13.2). Notice that to calculate $\hat{\omega}$ you must first translate frequencies to proportions. For the data in Table 20.1:

BRAND	p_o	p_e	$p_o - p_e$	$\frac{(p_o - p_e)^2}{p_e}$
A	15/100 = .15	.25	−.10	.0400
B	36/100 = .36	.25	.11	.0484
C	28/100 = .28	.25	.03	.0036
D	21/100 = .21	.25	−.04	.0064
			0	.0984

$$\hat{\omega} = \sqrt{.0984} = .31$$

Cohen (1992) suggests that we may interpret effect size, $\hat{\omega}$, as follows: .1 = small effect, .3 = medium effect, and .5 = large effect. Thus, in our soft drink problem, our outcome has a medium effect size.

20.7 Assumptions in the Use of the Theoretical Distribution of Chi-Square

Many people refer to chi-square as a nonparametric test, but it does, in fact, assume the central limit theorem and is therefore neither a "distribution-free" nor an "assumption-free" procedure (Bradley, 1968). The distribution of χ^2 follows from certain theoretical considerations, and thus we may expect obtained values of χ^2 to follow this distribution exactly only when the assumptions appropriate to this model are satisfied. What are these assumptions?

1. It is assumed that the sample drawn is a random sample from the population about which inference is to be made. In practice, this requirement is seldom fully met (see Section 13.11).
2. It is assumed that observations are independent. For example, in the soft drink experiment with 100 randomly selected subjects, the preference of one individual is not predictable from that of any other. But suppose each subject had been given three trials. The three responses of each subject are not independent of each other. In general, *the set of observations will not be completely independent when their number exceeds the number of subjects.*
3. It is assumed that, in repeated experiments, the observed frequencies will be normally distributed about the expected frequencies. With random sampling, this tends to be true.

20.8 Chi-Square as a Test for Independence between Two Variables

bivariate distribution a distribution that shows the relation between two variables

marginal totals row totals and column totals in a bivariate distribution

So far, we have considered the use of chi-square only in the one-variable case. However, we can also use it to analyze **bivariate frequency distributions**. For example, suppose we poll 400 students on a college campus and ask what role they feel students should have in determining the curriculum. We wish to analyze their responses in relation to their majors. Do students with different majors see their role in the same way?

To study this question, we classify the data (observed frequencies) in a bivariate distribution, as shown in Table 20.3. For each variable, the categories are mutually exclusive. What does Table 20.3 tell us? Note that of the 400 students, 200 were majors in humanities and arts (total frequency for the first row). We also see that 100 of the students wanted voting membership on the curriculum committee (total frequency for the third column). These row and column totals are often referred to as **marginal totals** because they are found at the margins of the table. The table also gives more specific information. For

TABLE 20.3 *Classification of Responses of 400 College Students to the Question: "How Should Students Participate in Determining the Curriculum?"*

		Desired Role of Students			
		Recommendation through student association	*Nonvoting membership on curriculum committee*	*Voting membership on curriculum committee*	f_{row}
Major of respondents	*Humanities and arts*	46	82	72	200
	Science	42	38	20	100
	Business	52	40	8	100
	f_{col}	140	160	100	$n = 400$

example, of the 200 humanities and arts majors, 72 thought that students should be voting members of the college curriculum committee. The frequencies in each cell (box) are called **cell frequencies**.

cell frequency the frequency of observations in a cell of a bivariate distribution

In many ways, this table is similar to the bivariate frequency distributions you encountered in the study of correlation (see Chapter 7). Indeed, the major difference is that the two variables (student's major and student's role) are both qualitative variables rather than quantitative variables. As with a correlation table, we may ask whether there is a relationship between the two variables in the population sampled. That is, is classification of response independent of classification of the respondent's major, or is one contingent on the other in some way?

contingency tables a bivariate distribution of two qualitative variables

Bivariate frequency distributions of the type illustrated in Table 20.3 are known as **contingency tables**. From such a table we can determine what cell frequencies would be expected if the two variables are independent of each other *in the population*. Then we may use *chi-square to compare the observed cell frequencies with those expected under the null hypothesis of independence.* If the $(f_o - f_e)$ differences are small, χ^2 will be small, suggesting that the two variables could be independent. Conversely, a large χ^2 points toward a contingency relationship.

***H_0* for chi-square test for independence between two variables** states that in the population, there is no relationship between the two variables; in practice, a test about proportions

In general, the **null hypothesis of independence** for a contingency table is equivalent to hypothesizing that in the population, the *relative* frequencies for any row (across the categories of the column variable) are the same for all rows, or that in the population, the *relative* frequencies for any column (across the categories of the row variable) are the same for all columns. So again we can think of the null hypothesis to be tested by χ^2 as one concerning proportions. For example, look at the data of Table 20.3. There are twice as many humanities and arts majors as there are science or business majors, but if the role students assign to themselves is unrelated to their major, then *on a proportional basis* the role assigned should be the same for students from each of the three majors. To calculate χ^2, we will begin by calculating the sample marginal proportions for columns—the proportion of cases appearing in each column when the three student majors are combined. We do this by dividing each column total by the grand total (n):

$$p_1 = \frac{\begin{bmatrix}\textit{number of students}\\ \textit{for recommendation}\\ \textit{through student}\\ \textit{association}\end{bmatrix}}{\textit{total number of students}} = \frac{f_{\text{col }1}}{n} = \frac{140}{400} = .35$$

$$p_2 = \frac{\begin{bmatrix}\textit{number of students}\\ \textit{for nonvoting membership}\\ \textit{on college curriculum}\\ \textit{committee}\end{bmatrix}}{\textit{total number of students}} = \frac{f_{\text{col }2}}{n} = \frac{160}{400} = .40$$

$$p_3 = \frac{\begin{bmatrix}\textit{number of students}\\ \textit{for voting membership}\\ \textit{on college curriculum}\\ \textit{committee}\end{bmatrix}}{\textit{total number of students}} = \frac{f_{\text{col }3}}{n} = \frac{100}{400} = .25$$

In the absence of a dependent relationship, we should expect the distribution of proportions *in each row* to approximate these figures. They are therefore the *expected* cell proportions, as shown in Table 20.4.

Table 20.4 *Expected Cell Proportions for Each Row in a Contingency Table (Data from Table 20.3)*

		Desired Role of Student		
		Recommendation through student association	*Nonvoting membership on curriculum committee*	*Voting membership on curriculum committee*
Major of respondents	*Humanities and arts*	$p_e = .35$	$p_e = .40$	$p_e = .25$
	Science	$p_e = .35$	$p_e = .40$	$p_e = .25$
	Business	$p_e = .35$	$p_e = .40$	$p_e = .25$
	Total	$p = \frac{140}{400} = .35$	$p = \frac{160}{400} = .40$	$p = \frac{100}{400} = .25$

20.9 Finding Expected Frequencies in a Contingency Table

Under the null hypothesis of independence of classification of the two variables in the population, we can easily calculate the expected cell frequencies from the expected cell proportions: In each row, we simply *multiply the expected proportion for each cell by the total number of subjects in that row.* For example, for the cell representing humanities and arts majors who thought that students should be voting members on the college curriculum committee, the expected frequency is:

$$\begin{aligned} f_e &= \left[\begin{array}{c}\textit{expected proportion}\\ \textit{for voting membership}\end{array}\right] \times \left[\begin{array}{c}\textit{total number of}\\ \textit{humanities and arts majors}\end{array}\right] \\ &= (.25)(f_{\text{row1}}) \\ &= (.25)(200) = 50 \end{aligned}$$

We show the complete results for our student example in Table 20.5.

Table 20.5 *Expected Frequencies for Each Cell in a Contingency Table (Data from Table 20.3)*

		Desired Role of Student			
		Recommendation through student association	*Nonvoting membership on curriculum committee*	*Voting membership on curriculum committee*	f_{row}
Major of respondents	*Humanities and arts*	$f_e = .35 \times 200 = 70$	$f_e = .40 \times 200 = 80$	$f_e = .25 \times 200 = 50$	200
	Science	$f_e = .35 \times 100 = 35$	$f_e = .40 \times 100 = 40$	$f_e = .25 \times 100 = 25$	100
	Business	$f_e = .35 \times 100 = 35$	$f_e = .40 \times 100 = 40$	$f_e = .25 \times 100 = 25$	100
		$f_{\text{col}} = 140$	$f_{\text{col}} = 160$	$f_{\text{col}} = 100$	$n = 400$
		$p_{\text{col}} = \frac{140}{400} = .35$	$p_{\text{col}} = \frac{160}{400} = .40$	$p_{\text{col}} = \frac{100}{400} = .25$	

To sum up, we may calculate frequencies expected under the null hypothesis of independence in a contingency table as follows:

Step 1: Find the column proportions by dividing each column total (f_{col}) by the grand total (n). In our example, they are .35, .40, and .25, respectively. The sum of these proportions should always be unity.

Step 2: Multiply each row total by these column proportions; the result in each instance is the expected cell frequency (f_{e}) for the cells in that row. This is shown in the cells of Table 20.5. Thus, we calculate the expected frequency for any cell by

$$f_{\text{e}} = \left(\frac{f_{\text{col}}}{n}\right)(f_{\text{row}})$$

We could also have obtained the same results by finding the row proportions and multiplying them by the column totals:

$$f_{\text{e}} = \left(\frac{f_{\text{row}}}{n}\right)(f_{\text{col}})$$

For comprehension purposes, it is best to learn to calculate expected frequencies by column proportions or row proportions, as we have done. However, there is a shortcut. Look at the previous two formulas. Either one may easily be transformed to

$$\boldsymbol{f_{\text{e}} = \frac{(f_{\text{row}})(f_{\text{col}})}{n}}$$

shortcut formula for calculating expected cell frequencies

$$f_{\text{e}} = \frac{(f_{\text{row}})(f_{\text{col}})}{n}$$

As an example, for the cell representing humanities and arts majors who thought that students should be voting members on the college curriculum committee,

$$f_{\text{e}} = \frac{(200)(100)}{400} = 50$$

Step 3: Check to see that the total of the expected frequencies in any row or in any column equals that of the observed frequencies. If it does not, you have made an error in calculation.

20.10 Calculation of χ^2 and Determination of Significance in a Contingency Table

Now that we have the observed frequencies (Table 20.3) and expected frequencies (Table 20.5) for each cell, we calculate χ^2 according to Formula 20.1. As usual, each ($f_{\text{o}} - f_{\text{e}}$) difference is squared and divided by f_{e}. The sum of these nine components is the calculated value of χ^2:

$$\begin{aligned}\chi^2 &= \frac{(46-70)^2}{70} + \frac{(42-35)^2}{35} + \frac{(52-35)^2}{35} + \frac{(82-80)^2}{80} + \frac{(38-40)^2}{40} \\ &\quad + \frac{(40-40)^2}{40} + \frac{(72-50)^2}{50} + \frac{(20-25)^2}{25} + \frac{(8-25)^2}{25} \\ &= 40.28\end{aligned}$$

How many degrees of freedom are associated with this χ^2? To figure *degrees of freedom in a contingency table,* we consider that the row totals and the column totals are fixed

R
symbol for number of rows

C
symbol for number of columns

$(R - 1)(C - 1)$
degrees of freedom in a contingency table

and ask how many cell frequencies are free to vary. In general, if ***R*** is the number of rows and ***C*** is the number of columns, the number of cell frequencies that can vary (and therefore the number of degrees of freedom) for an $R \times C$ contingency table is $(R - 1)(C - 1)$. For our problem, $df = (3 - 1)(3 - 1) = 4$.

If the null hypothesis of independence is true at the population level, random sampling would result in obtained values of χ^2 that are close to the tabulated distribution of that statistic. If the null hypothesis is false in any way, the obtained values of χ^2 will tend to be larger than when the hypothesis is true. As before, then, the region of rejection is placed in the upper tail of the tabulated distribution. For $df = 4$, Table I in Appendix D shows that when the test is conducted at $\alpha = .05$, $\chi^2_{\text{crit}} = 9.49$, and when $\alpha = .01$, $\chi^2_{\text{crit}} = 13.28$. The calculated χ^2 is 40.28, which exceeds both critical points. Accordingly, we reject the null hypothesis of independence.

Remember that a significant outcome of the chi-square test is directly applicable *only to the data taken as a whole.* The χ^2 that we obtained is inseparably a function of the nine contributions (one from each cell) composing it. We cannot say for sure whether one group is responsible for the finding of significance or whether all are involved.

20.11 Measures of Effect Size (Strength of Association) for Tests of Independence

ϕ
symbol for the phi coefficient and for Cramer's phi, two measures of the strength of association

$\sqrt{\dfrac{\chi^2}{n}}$
formula for the phi coefficient, for a 2 × 2 contingency table

Because the contingency table is analogous to the correlation table, you might think that χ^2, like r, provides a measure of strength of association. However, χ^2 does not by itself serve this function. The only purpose of the chi-square test as applied to a contingency table is to test a null hypothesis of independence between two variables. But what if the outcome of our χ^2 test leads us to reject the null hypothesis? That means there is some degree of correlation between the variables, but how strong is it? Two measures of association are commonly used *after* the null hypothesis of independence has been rejected. Both are members of the r family of effect sizes (see Section 14.9).

In the case of a 2 × 2 contingency table, a measure of strength of association is provided by the **phi coefficient (ϕ)**:

FORMULA FOR THE PHI COEFFICIENT

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{20.3}$$

where: χ^2 = the calculated χ^2 value
n = sample size.

The value of ϕ ranges from 0 (no association) to 1 (perfect association).[1] J. Cohen (1988) suggests that we may interpret ϕ in the same manner as we did $\hat{\omega}$: .1 = small effect, .3 = medium effect, and .5 = large effect.

We may use phi only with a 2 × 2 table. For larger contingency tables, the appropriate measure of association is **Cramer's phi** (also ϕ, although sometimes symbolized as **V**):

---

[1] In a 2 × 2 table where we have two dichotomous variables, if we order the categories for each variable and assign a score of 0 to the lower category and 1 to the higher category, ϕ is equivalent to Pearson's r.

Point of Controversy

Yates' Correction for Continuity

Theoretical chi-square distributions are smooth and continuous. Calculated chi-square statistics, on the other hand, have only discrete values and are discontinuous (with irregular, stepwise distributions). Thus, differences may exist between calculated and theoretical values of χ^2. To compensate for this, Yates (1934) introduced a **correction for continuity,** which cons ists of reducing the differences between f_o and f_e by .5 before squaring:

$$\chi^2 = \frac{(|f_o - f_e| - .5)^2}{f_e}$$

Yates' procedure reduces the calculated value of χ^2, thereby making a significant result less likely.

Older statistics textbooks, including previous editions of this text, almost always recommended use of Yates' correction in cases where $df = 1$ (including 2 × 2 contingency tables), particularly if the expected frequency in one of the cells is less than 5. Statisticians have now shown, however, that *the uncorrected chi-square statistic is reasonably accurate when the average expected frequency is as low as 2* (e.g., Camilli & Hopkins, 1978; Conover, 1974; Delucchi, 1983; Roscoe & Byers, 1971). Consequently, the correction has been abandoned for the one-variable (goodness-of-fit) case. With regard to 2 × 2 tables, old habits linger on, and many investigators continue to use Yates' correction.

There are, in fact, some statisticians who still favor the use of Yates' correction in a 2 × 2 table (Fleiss, 1981; Greenhouse, 1990; Mantel, 1990; Mantel & Greenhouse, 1968). Most statisticians disagree. *One* of the problems is that Yates developed his procedure for situations in which the experimenter controls both sets of marginal frequencies in the 2 × 2 table, but this is rarely encountered in practice. When applied to most studies (where marginal frequencies are free to vary), Yates' correction makes an already very conservative test even more conservative, "causing the actual proportion of Type I errors to be substantially lower than the nominal alpha levels" (Haviland, 1990). The result is greatly reduced power. Computer simulation studies have shown the uncorrected chi-square test to work quite well for tests of independence (Camilli & Hopkins, 1978).

A word of caution: The use of small expected frequencies increases the chance of making a Type II error (failure to reject a false null hypothesis). Thus, you would still be wise to use a large sample size. For an excellent review of this subject, including comments from both those who favor and those who disfavor Yates' correction, see Haviland (1990).

$$\sqrt{\frac{\chi^2}{n(df_{\text{smaller}})}}$$

formula for Cramer's phi, for use with any contingency table

FORMULA FOR CRAMER'S PHI

$$\phi = \sqrt{\frac{\chi^2}{n(df_{\text{smaller}})}} \tag{20.4}$$

where: χ^2 = the calculated χ^2 value
n = sample size
df_{smaller} = df for the *smaller* of the number of rows and columns.

Notice that Cramer's phi is really just an extension of the phi coefficient for a 2 × 2 table. In the case of a 2 × 2 table, df_{smaller} in Formula 20.4 is 1, and phi = Cramer's phi. Thus, we may use Cramer's phi for any contingency table.

As an example, look again at Table 20.3. Recall that our obtained χ^2 was 40.28 and we rejected the null hypothesis of independence. There were three rows and three columns, so $df_{\text{smaller}} = 2$. We calculate Cramer's phi as

$$\phi = \sqrt{\frac{40.28}{(400)(2)}} = \sqrt{.05035} = .22$$

The phi coefficient is equivalent to $\hat{\omega}$, a measure of effect size that we introduced in Section 20.6. However, Cramer's phi is not equivalent to $\hat{\omega}$. To convert Cramer's phi to $\hat{\omega}$, multiply it by the square root of df_{smaller}:

FORMULA FOR CONVERSION
OF CRAMER'S PHI TO $\hat{\omega}$

$$\hat{\omega} = \text{Cramer's } \phi\sqrt{df_{\text{smaller}}} \qquad (20.5)$$

As before, J. Cohen (1988) states that .1 = small effect, .3 = medium effect, and .5 = large effect.

Another measure of association often used in the case of the 2 × 2 contingency table is the **odds ratio**. We briefly mentioned it in Section 10.6. The odds ratio can range between 0 and ∞. When the odds ratio equals 1, it indicates that the two variables are independent. As the odds ratio deviates from 1, this suggests dependence between the two variables.

Before computing the odds ratio, we could compute the odds first using the observed frequencies in a 2 × 2 contingency table. In Table 20.6, for example, 137 job applicants were classified in terms of their sex and hiring status. In the table, we see that 70 job applicants were hired and, of these, 20 were female. Using Formula 10.2, we can compute the odds of being hired for females, which is $\Omega_{\text{Hired (females)}} = (20/65)/[1-(20/65)] = .444$, and the odds of being hired for males, which is $\Omega_{\text{Hired (males)}} = (50/72)/[1-(50/72)] = 2.273$. Now, with these two odds, the population odds ratio (θ) can be estimated using $OR = \Omega_{\text{Hired (females)}} / \Omega_{\text{Hired (males)}} = .1956$. Here, we use OR to represent an estimate of θ. A shortcut formula to calculate OR without computing the odds first would be to use the following:

$$OR = \frac{AD}{BC} \qquad (20.6)$$

where A, B, C, and D simply correspond to the observed frequencies in each cell (see Table 20.6). Thus, $OR = (20 \times 22) / (45 \times 50) = .1956$.

Notice that if the odds of being hired were exactly the same for females and males, OR would equal 1. This would indicate that sex and hiring status are not related. However, as OR deviates from 1, this suggests dependence. More specifically, if $OR > 1$, this indicates

TABLE 20.6 *Classification of 137 Job Applicants Based on Sex and Hiring Status*

		Hiring Status		
		Hired	Not Hired	f_{row}
Sex	Female	20 A	45 B	65
	Male	50 C	22 D	72
	f_{col}	70	67	$n = 137$

TABLE 20.7 *Power for the Chi-Square Test of Independence, $\alpha = .05$*

df	*n*	SMALL EFFECT	MEDIUM EFFECT	LARGE EFFECT
1	50	.11	.56	.94
	100	.17	.85	.99+
	200	.29	.99	.99+
2	50	.09	.46	.90
	100	.13	.77	.99+
	200	.23	.97	.99+
3	50	.08	.40	.86
	100	.12	.71	.99
	200	.19	.96	.99+
4	50	.08	.36	.82
	100	.11	.66	.99
	200	.17	.94	.99+

Source: Data for this figure were obtained from J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed., Lawrence Erlbaum Associates, Inc., New York, 1988, and are used with permission.

that the odds in the first row of the contingency table are *greater* than the odds in the second row. However, if $OR < 1$, this indicates that the odds in the first row of the contingency table are *less* than the odds in the second row. Note that an *OR* of 8 is farther from independence than an *OR* of 2. Similarly, an *OR* of .125 is farther from independence than an *OR* of .5. Based on Table 20.6, the odds of being hired for females are less than those for males. In particular, the odds of being hired for females are .1956 those for males.

We could have switched the rows in Table 20.6 so that males appear in the first row and females in the second row. This *OR* would simply be the reciprocal of .1956. That is, $1/.1956 = 5.11$. This indicates that the odds of being hired for males are slightly more than 5 times those for females. Thus, if we switch the order of the rows (or columns), the new *OR* is simply the reciprocal of the original value. This is convenient and as we noted in Section 10.6, the *OR* is very popular and is used in more advanced statistical procedures because of its flexibility (Agresti, 2002; Hosmer & Lemeshow, 2000).

20.12 Power and the Chi-Square Test of Independence

What about power? J. Cohen (1988) has computed tables based on the effect size $\hat{\omega}$. We reproduce a small part of Cohen's tables in Table 20.7. As an example, suppose you planned a study with a 2×3 contingency table using 100 subjects. If you anticipate a medium effect size, what would be the power of your test when $\alpha = .05$? For $df = 2$ and $n = 100$, Table 20.7 shows us that power would be approximately .77. Therefore, if the null hypothesis is false (the two variables are really not independent), you have about a 77% chance of rejecting the hypothesis. For more complete tables, see Cohen's text.

20.13 Summary

In previous chapters, we introduced inferential techniques that deal with quantitative variables and are used for testing hypotheses about the population mean or the value of the correlation coefficient. Our concern in this chapter has been inferences about frequency—the number of cases falling into the two or more categories of a population distribution. The statistical test we used to make inferences about frequencies is called chi-square (χ^2).

In tests of frequency in categories of a single variable (often called tests for goodness of fit), the chi-square test tells us whether the relative frequencies observed in the two or more categories of our sample frequency distribution are in accord with the set of such values hypothesized to be true in the population. The test determines whether the differences between the observed frequencies and the frequencies expected under H_0 are greater than would be anticipated on the basis of chance (due to random sampling variation). When H_0 is true, the sample χ^2s follow a theoretical distribution with $k - 1$ degrees of freedom, where k is the number of categories. Larger differences between f_o and f_e result in larger calculated values of χ^2, so the region of rejection is always in the upper tail of the sampling distribution. With just two categories, we can test a hypothesis about a single proportion, but with more than two categories, the alternative hypothesis is very broad (H_0 can be untrue in many ways) and any conclusion must be taken as a tentative hypothesis. A measure of effect size is provided by $\hat{\omega}$.

The chi-square test may also be used in the analysis of bivariate frequency distributions to test whether the two variables are independent. To conduct the test, we cross-tabulate the data in a contingency table and calculate the $(f_o - f_e)$ difference for each cell. If our calculated χ^2 is larger than the critical value for $(R - 1)(C - 1)$ degrees of freedom, we reject the null hypothesis and conclude that the two variables are related in *some* way. When the null hypothesis of independence is rejected, that means there is some degree of association between the two variables. The strength of association is provided by the phi coefficient (ϕ) for 2 × 2 tables and by Cramer's phi (ϕ) for larger tables. In addition, the odds ratio (OR) is a common measure of association.

Key Terms, Concepts, and Symbols

categorical data (p. 364)
P (p. 364)
goodness-of-fit problems (p. 364)
k (p. 364)
observed frequencies f_o (p. 365)
expected frequencies f_e (p. 365)
chi-square statistic (p. 365)
χ^2 (p. 366)
null and alternative hypotheses in the one-variable case (p. 364)
degrees of freedom (p. 366)
$k - 1$ (p. 366)
chi-square sampling distributions (p. 367)
$\hat{\omega}$ (p. 369)
bivariate distribution (p. 370)
marginal totals (p. 370)
cell frequency (p. 371)
contingency tables (p. 371)
null hypothesis of independence (p. 371)
expected frequencies in a contingency table (p. 372)
$(R - 1)(C - 1)$ (p. 374)
phi coefficient ϕ (p. 374)
Yates' correction for continuity (p. 375)
Cramer's phi (p. 376)
odds ratio (p. 376)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. In the χ^2 test, why is only the area in the upper tail of the χ^2 distribution of interest?

2. An investigator wishes to study the frequency of accidents among four divisions of an industry (H_0: $P_1 = P_2 = P_3 = P_4$). Is χ^2 appropriate for this study? Explain. Suppose the investigator wishes to study the frequency of accidents for all workers in the months of January, February, March, and April. Is χ^2 appropriate for this study? Explain.

3. A graduate student wants to know about a particular rat's preference for light or dark at two alternative routes at a choice point in a maze. Would it be permissible to give it 50 trials and use the χ^2 test? Explain.

4. We wish to determine whether a gambler's die is loaded. We roll the die 60 times and obtain the following results:

Side coming up:	1	2	3	4	5	6
Number observed:	8	8	5	10	14	15

(a) In Problem 3, one rat was given 50 trials; in this problem one die is given 60 trials. Why was it not appropriate to use the χ^2 test in Problem 3, whereas it is appropriate here? (See the second assumption in Section 20.7.) (b) State, in symbolic form, two equivalent statements of H_0 for this situation. (c) Can H_A be written in a single symbolic statement? Explain. (d) Compute the expected frequencies under H_0. (e) Compute χ^2 and complete the test ($\alpha = .05$). Draw final conclusions. (f) Do these results prove the die is not loaded? (See Section 12.6.)

5. In a grocery store, customers who pass by are asked to sample three cheese spreads and to declare their preference for one. Their frequencies of choice are as follows:

Spread:	A	B	C
Frequency of choice:	20	39	25

(a) Test the null hypothesis of equal preference using $\alpha = .05$ and .01. What conclusions are allowable when χ^2 proves to be significant? (b) What is the effect size, $\hat{\omega}$, for this result? (c) Is this a small, medium, or large effect?

6. For Problem 5, how does the method of sampling affect the researcher's ability to make inferences about the population? (See Section 13.12.)

7. One hundred students take a multiple-choice test. We wish to test the hypothesis that the students guessed at random on question 36. The frequencies of responses for that question were as follows:

Alternative:	A	B	C	D	E
Frequency of response:	15	40	5	12	28

(a) State H_0 in symbolic form. (b) Compute χ^2 and test H_0 at $\alpha = .01$. (c) Draw final conclusions. (d) Calculate $\hat{\omega}$ for this result and state whether this is a small, medium, or large effect size.

8. Refer to Problem 7. Suppose alternative B is the correct answer. (a) For this alternative, state the observed frequencies of correct and incorrect answers. (b) What are the expected frequencies if the students are just guessing? (c) Suppose we wish to test the null hypothesis that the frequencies of correct and incorrect answers to alternative B are in accord with random guessing. State H_0 in terms of the proportion guessing correctly. (d) Compute χ^2 and perform a nondirectional test of H_0 at $\alpha = .05$. Draw final conclusions. (e) Suppose we wish to test H_0 against the alternative that at least some students know the correct answer. State H_A. (f) Why can we perform a directional test here but not in Problem 7?

9. You may have heard the old adage, "Stay with your first answer on a multiple-choice test." Is changing answers more likely to be helpful or harmful? To examine this, Best (1979) studied the responses of 261 students in an introductory psychology course. He recorded the number of right-to-wrong, wrong-to-right, and wrong-to-wrong answer changes for each student. More wrong-to-right changes than right-to-wrong changes were made by 195 of the students, who were thus "helped" by changing answers; 27 students made more right-to-wrong changes than wrong-to-right changes and thus "hurt" themselves. (a) For these 222 students, conduct a χ^2 test of the hypothesis that the population proportions of right-to-wrong and wrong-to-right changes are equal. Your conclusions? (b) What is the effect size?

10. A botanist plants a group of hybrid flower seeds she has developed. On the basis of a certain genetic theory, she predicts that, on the average, 1/4 of the resulting flowers will be red, 1/4 will be blue, and 1/2 will be purple. The following are the results for the seeds she has planted:

COLOR OF FLOWER		
Red	***Blue***	***Purple***
19	38	63

(a) State H_0. (b) Compute χ^2 and test H_0 at $\alpha = .05$. Do the results tend to support or disprove the theory as applied to this type of seed? (c) Calculate $\hat{\omega}$.

11. When polled about a proposition on a ballot, the responses of men and women were as follows:

	AGREE	UNDECIDED	DISAGREE
MEN	70	25	40
WOMEN	30	20	25

(a) State in words the null hypothesis of independence for this situation. (b) Complete the χ^2 test ($\alpha = .05$) and draw a conclusion. (c) Convert each observed frequency into a proportion based on its row frequency. What interpretation seems likely?

12. The mechanical aptitude scores of different classifications of workers are as follows:

	30–49	50–69	70 AND ABOVE
UNSKILLED	20	30	10
SEMISKILLED	10	40	30
SKILLED	5	10	20

(a) Test the null hypothesis that test score is independent of job classification. Use $\alpha = .05$ and $\alpha = .01$. (b) Translate each observed frequency into a proportion relative to its row frequency. Do the same for the expected frequencies. Comparing the two, and in view of your answer to (a), what interpretation appears likely? (c) Calculate the strength of association, ϕ. (d) Calculate effect size, $\hat{\omega}$.

13. Fifty rats are randomly divided into Group A and Group B. The rats in Group A are given a series of training trials that differ in several respects from those given to Group B. Then each rat is given a test trial in which it has a choice of leaving the maze through a dark exit or a light exit. The following are the results for the test trial:

	EXIT SELECTED	
	Dark	*Light*
Group A	7	18
Group B	12	13

You wish to determine whether rats trained like those in Group A differ from Group B rats in their exit preferences. (a) Give H_0 and H_A in terms of the proportions preferring the dark exit (use symbols). (b) Compute χ^2. Use $\alpha = .05$. (c) Draw final conclusions. (d) Calculate the phi coefficient and odds ratio.

14. Forty subjects volunteer for an experiment on attitude change. An attitude item is administered to the subjects before they see a particular TV movie and then again after they see the movie, with the following results:

	RESPONSE TO ATTITUDE STATEMENT		
	Agree	*Undecided*	*Disagree*
Before	8	20	12
After	18	12	10

Can we use the χ^2 test described in this chapter to test the hypothesis that time of administration (before or after) is independent of statement response? Explain.

15. A survey was taken to see whether there is a relationship between political affiliation and attitudes toward a bond issue that will be on an upcoming ballot. The following results were obtained:

	RESPONSE TO BOND ISSUE		
	Approve	*Disapprove*	*Undecided*
Republican	20	80	20
Democratic	90	20	30
Independent	20	5	15

(a) State the null hypothesis. (b) Compute χ^2; use $\alpha = .05$. (c) State the conclusion. (d) Calculate the strength of association and effect size. (e) Why is it appropriate to use χ^2 for this problem, but not for Problem 14?

CHAPTER 21

Some (Almost) Assumption-Free Tests

When you have finished studying this chapter, you should be able to:

- Understand that parametric tests, such as t and F, make the assumption that the population(s) from which we draw our sample(s) is normally distributed;
- Explain why, when sample size is small, violations of this assumption may lead to incorrect conclusions when using t or F;
- Understand that assumption-freer tests (e.g., randomization tests, rank-order tests) do not make assumptions about the underlying population(s), and are more powerful than parametric tests when the population is not normally distributed and sample size is small; and
- Calculate the Mann-Whitney U test, the sign test, the Wilcoxon signed-ranks test, the Kruskal-Wallis test, and Friedman's rank test.

The inferential statistical techniques that we have considered up to this point—t, F, and χ^2—are widely used, but they require assumptions about population distributions that may be incorrect in a lot of cases. For example, the t and F tests of the difference between (among) means of independent samples assume normality of the group populations and homogeneity of variance. However, as you have learned, many real-world variables are distributed in a manner that differs substantially from normality (Micceri, 1989; see Point of Controversy in Chapter 6). Moderate departures from normality (and even substantial departures from homogeneity of variance) may not seriously invalidate these tests when sample size is moderate to large, but the problem of violation of assumptions is of great concern when sample size is small (<25). Why?

Recall from Section 11.4 that even when the population distribution greatly departs from normality, the central limit theorem states that the sampling distribution of $\overline{X}$ will nevertheless approach a normal distribution as sample size increases (and will resemble a normal curve when $n = 25$). Unless the population distribution is normal in shape, the use of samples with $n < 25$ is risky, and is more so the greater the population deviates from normality. Note that you can use IBM SPSS to check assumptions of normality and homogeneity of variance.

In this chapter, we introduce some alternative procedures for testing hypotheses that make less restrictive assumptions about the population distributions. These procedures are often referred to as **nonparametric statistics** because the statistical hypotheses that are tested are not about parameters such as μ, or as **distribution-free methods** if the procedure does not make any assumptions about the shape of the population. These two terms are often used interchangeably, but they do have different meanings. Consequently, Ury (1967) suggested an alternative, more descriptive name—**assumption-freer tests**. "Freer" emphasizes that the tests are not completely free of assumptions; it is just that the assumptions are less restrictive than for t, F, or χ^2. (Many people call chi-square a nonparametric test, but it does in fact assume the central limit theorem; see Bradley, 1968.)

When the assumption of normality is met, the parametric tests (e.g., t and F) are more powerful than the assumption-freer tests. However, when the assumptions are *not* met (which is often the case), the assumption-freer methods are more powerful. That is, they are more likely than t or F to give a significant result when the null hypothesis is false.

nonparametric statistics tests that do not test hypotheses about population parameters

distribution-free methods tests that do not make any assumptions about the shape of the population(s) from which the sample(s) is (are) drawn

assumption-freer tests tests that make fewer assumptions than the traditional parametric tests such as t or F

21.1 The Null Hypothesis in Assumption-Freer Tests

The t and F tests focus on the mean of a distribution or the difference between means of two (or more) distributions. The techniques described here are less specific. *The usual hypothesis is simply that the distributions are not different in any respect,* rather than that there is no difference among the distributions with regard to the mean. This suggests (correctly) that "significance" could result from some unknown combination of differences in central tendency, variability, and shape. However, if distributions are not of greatly differing shape, most are good tests of central tendency (even though they do not specifically refer to a particular measure of central tendency) and may, in fact, be the preferred procedures with small sample sizes.

21.2 Randomization Tests

Our text follows a decades-old tradition in its focus on inferential techniques based on the statistics t, F, and χ^2. One of the assumptions of these tests is that each sample of observations under analysis comes from a normally distributed population through sampling at random. In many experiments, however, the assumption of random sampling is not met. Researchers typically obtain subjects through an informal process, taking whoever is convenient. They usually introduce randomness into their investigation only in choosing which condition a subject will serve in.

In an experiment investigating the effect of aerobic exercise on depression, for example, we might enroll the next 30 female students who come to our college's health center for treatment of depression. We might then flip a coin to determine whether each participant will be subjected to the experimental treatment (a thrice-weekly aerobics class, say) or simply put on a waiting list for therapy (design from McCann & Holmes, 1984). This process is called *random assignment* or *randomization* (see Sections 14.12 and 14.13).

Statisticians have developed tests that are free of the assumptions of normality and homogeneity of variance in the parent population, and most do not require random sampling (May, Masson, & Hunter, 1997). They do require that scores or pairs of scores be independent of each other, but this is achieved by random assignment rather than by random sampling. **Randomization tests** directly address the question of whether the batches of observations differ only because of chance. The essential insight from which randomization tests derive is Sir Ronald Fisher's (1935).

randomization tests
tests that use random assignment and compare obtained results with all possible combinations of scores

To conduct a randomization test for an experiment like the one on aerobics and depression, we start with the observations on hand, collecting them into one bunch. We then assign each observation at random to one of the two conditions. Think of the conditions as a pair of boxes resting side by side in front of us. In effect, we are tossing each score at the boxes blindly so that the scores have an equal chance of landing in each box. When all scores are assigned (randomly) between the two boxes, we find the mean for each box and then the difference between the means.

Now we repeat the process. From the undifferentiated batch of scores—the actual scores that we obtained when measuring the subjects for depression—we again pull them out, one at a time (the order doesn't matter) and toss them blindly at those boxes. Again we summarize the result as the difference between the means of the two batches so created. Because of the randomness, it is unlikely that we will find the same difference that occurred the first time.

We continue to repeat the process until we have generated *all possible combinations of scores*. With many scores in each group, this may mean thousands, perhaps millions, of different combinations. Such calculations were not possible before computers, but the modern computer can do this work for us quickly. What we now have is an *actual* (not theoretical) distribution of the differences between means, each difference the result of the randomization of the scores *of our particular experiment*. We know this distribution through observation; we generated it, and we can directly look at it.

As a simple example, suppose we have six subjects that we randomly assign to one of two groups of three, and that when the study is over the subjects have scores of 1, 2, 3, 4, 5, and 6. How many ways could the study have turned out? Look at Table 21.1. It shows that there are 20 possible outcomes (determined by $6!/(6 - 3)!3!$). However, when we compare the difference between the means of the two groups, we see that there are 10 possible outcomes, some more probable than others.

Now we simply compare the difference between the two means that turned up in our experiment with the distribution we generated. Suppose our difference proves to be small compared to most in the distribution, say, -0.33 or $+0.33$. This is a quite ordinary difference, one that was not included among the 5% most extreme outcomes. It is therefore plausible that the difference is due only to randomization, so we have no grounds for concluding that anything systematic caused the scores in the two conditions to differ.

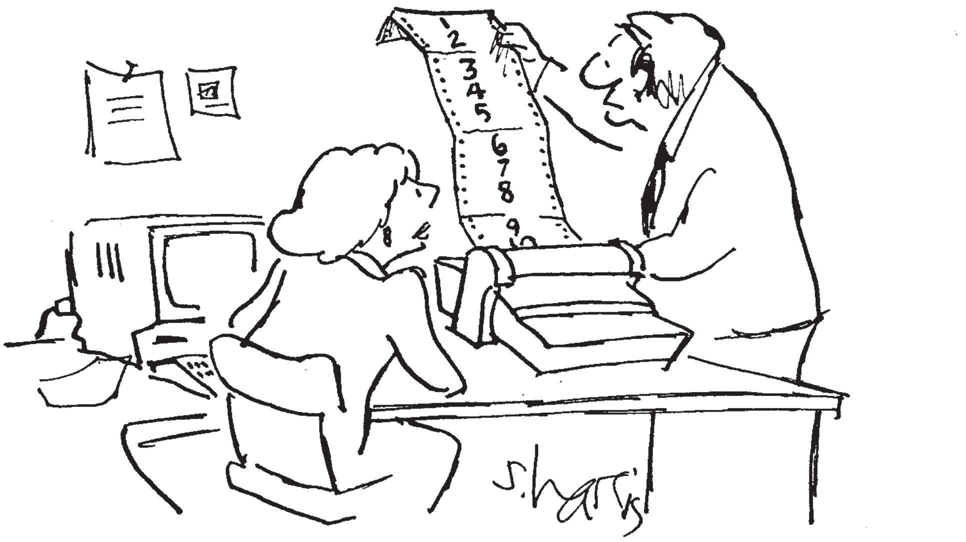
Do you recognize the logic of hypothesis testing here? The null hypothesis says that our subjects' depression scores are unaffected by which condition of the experiment they served in. According to this null hypothesis, the two batches of scores differ only because of randomization. Because our obtained difference between the means is one of those that would be common when only randomization is at work, we retain the null hypothesis.

But perhaps the actual difference will prove to be unusual when compared to the others that constitute our distribution of such quantities. Suppose the mean for the experimental condition is so much lower than the mean for the control condition that the difference is among the most rare 5%. Now it seems implausible that the difference between the two batches is due only to randomization. Now we are justified in rejecting the null hypothesis. We may

TABLE 21.1 *All Possible Outcomes of Random Assignment of Six Scores to Two Groups of Three*

OUTCOME	X	Y	$\overline{X} - \overline{Y}$
1	1, 2, 3	4, 5, 6	−3.00
2	1, 2, 4	3, 5, 6	−2.34
3	1, 3, 4	2, 5, 6	−1.66
4	1, 2, 5	3, 4, 6	−1.66
5	1, 2, 6	3, 4, 5	−1.00
6	1, 3, 5	2, 4, 6	−1.00
7	2, 3, 4	1, 5, 6	−1.00
8	1, 3, 6	2, 4, 5	−0.33
9	1, 4, 5	2, 3, 6	−0.33
10	2, 3, 5	1, 4, 6	−0.33
11	2, 3, 6	1, 4, 5	+0.33
12	2, 4, 5	1, 3, 6	+0.33
13	1, 4, 6	2, 3, 5	+0.33
14	1, 5, 6	2, 3, 4	+1.00
15	2, 4, 6	1, 3, 5	+1.00
16	3, 4, 5	1, 2, 6	+1.00
17	2, 5, 6	1, 3, 4	+1.66
18	3, 4, 6	1, 2, 5	+1.66
19	3, 5, 6	1, 2, 4	+2.34
20	4, 5, 6	1, 2, 3	+3.00

BIZARRE SEQUENCE OF COMPUTER-GENERATED RANDOM NUMBERS



© Sidney Harris, ScienceCartoonsPlus.com conclude that something systematic was at work. What's the risk that we are making an error? It's 5%—so we tested the null hypothesis at the .05 level of significance.

The familiar logic of hypothesis testing thus underlies a randomization test. We can employ a directional or a nondirectional alternative hypothesis, and we can choose any level of significance and thereby set our risk of a Type I error.

As the example shows, a randomization test leads directly to a conclusion of the kind an experimenter seeks: Is the difference between the batches of scores due only to chance or to something systematic? Had we drawn samples at random from the population of interest, then statistics would have also provided a basis for making generalizations to the population. However, when the groups are formed by random assignment of available subjects, the inferential statistics do not pretend to generalize beyond these scores. *Without random sampling (whether using parametric or assumption-freer tests), generalizations require judgment based on knowledge of the subject matter and on the similarity of new circumstances to those that characterize the previous studies of the topic* (see Sections 14.12 and 14.13).

21.3 Rank-Order Tests

rank-order tests tests in which original scores are first transformed into ranks

Most of the assumption-freer tests you will learn to use in this chapter are called **rank-order tests**. Original scores are transformed into ranks according to magnitude. Rank-order tests make no assumption about the shape of the distribution. In fact, if there are no ties in ranks, the distribution is known—rectangular (one score at each rank). Converting original scores to ranks has another advantage in that it effectively eliminates outliers. Recall that an outlier is an anomaly, a score that is very different from the others in the distribution. An outlier or two can often eliminate the possibility of obtaining a significant difference when using a parametric test (e.g., t or F).

After the scores are placed into ranks, we evaluate the outcome by the randomization procedures you learned in the previous section (Edgington, 1987). In essence, then, these tests are really *rank-order randomization tests*. You have already been introduced to one of them. Look again at Spearman's rank-order correlation coefficient from Chapter 7. The first four statistical tests that we discuss in this chapter were all derived from randomization tests. If you are interested in how these tests were derived, see Edgington (1987). For an introduction to many other randomization tests, see May, Masson, and Hunter (1997).

Most rank-order tests assume that the underlying variable is continuous. However, recall that many dependent variables in the behavioral sciences are discrete variables. This can result in identical scores, which will lead to ties in ranks. Ties in ranks will usually not cause a problem until one-third of the scores result in ties. Look at Section 7.9 to review how to assign ranks to tied scores.

21.4 The Bootstrap Method of Statistical Inference

The key problem in inferential statistics is how to infer the value of a population parameter from a sample drawn from that population. Conventional statistical methods, developed between 1800 and 1930, rely on mathematical equations, such as those for z, t, and F, that had to be calculated by hand. (The invention of calculators sped up the process somewhat.) As we've seen, these techniques compare an obtained sample with other samples that could be generated from the population (if the null hypothesis is true). To generate these samples, the methods generally assume that the population data are distributed normally.

bootstrap method
a testing technique that compares an obtained sample with (usually billions of) samples computer-generated from the original sample

In 1977, Bradley Efron, a statistician at Stanford University, invented a new method of statistical inference that takes advantage of high-speed computers (see Efron, 1979). In fact, the technique would not be possible without them. Efron's **bootstrap method** generates new samples from the original sample and requires no assumptions about the distribution of scores in the population.

To see how the bootstrap method works, let us suppose that we have randomly drawn a sample of 15 scores from a population of scores. Each of the 15 scores is copied by a computer a large number of times (a billion is not unreasonable with computers), and the 15 billion copies that result are thoroughly mixed. Thousands of additional samples, called *boot-strap samples*, of size 15 are then randomly sampled (with replacement) by the computer from this pool.

For each new sample, some of the 15 values might be drawn more than once and others not at all. The resulting range of values provides an index of the amount of variation in the statistic under examination (such as the mean). The statistical accuracy of the original sample statistic depends on the width of the interval associated with a particular percentage of all the samples. For example, if 95% of the calculated values of $\overline{X}$ are between 101 and 105, we would probably judge an original $\overline{X}$ of 103 to be highly accurate. If 95% of the $\overline{X}$s are between 88 and 115, on the other hand, we would place less faith in our original sample statistic. The method is called the bootstrap method because "the data, in a sense, pull themselves up by their own bootstraps by generating new data sets through which their reliability can be determined" (Kolata, 1988).

Overall, the bootstrap method provides an alternative way of calculating standard errors for any statistic. This can be especially handy for instances in which there exists no formula for calculating a standard error for a particular statistic.

How accurate is this new method? If the population from which the original sample is drawn is normally distributed, the bootstrap method and conventional methods produce the same answer. When the population deviates from normality, the bootstrap method is superior. This has been empirically demonstrated many times by drawing samples from populations with known characteristics. See Diaconis and Efron (1983) for some easy-to-understand examples.

It is probably safe to say that had inferential statistics been invented in this era of computers, its methods would be quite different from those presented in this text. New ideas take time to be accepted, but computer-intensive techniques may soon dominate statistics, and the techniques taught in this book may be relegated to a special box on "historical points of interest."

21.5 An Assumption-Freer Alternative to the *t* Test of a Difference between Two Independent Groups: The Mann-Whitney *U* Test

Mann-Whitney *U* test
a rank-order test of a difference between two independent groups

The **Mann-Whitney *U* test** is a popular alternative to the *t* test of the difference between means of two *independent* samples. We should consider using it when our measurements are weaker than interval scaling (see Section 1.5) or even with interval scale measurements when our samples are small and we have serious doubts about the distributional assumptions necessary for the *t* test.

The null hypothesis is that the two samples come from populations with the same *distribution*. If the two population distributions are of even moderately similar shape and variability, the Mann-Whitney *U* test is also an excellent test of central tendency. Because the test is on ranks, the most closely corresponding measure of central tendency is the median. Keep this in mind when evaluating the outcome of the test. (See Section 4.6 for properties of the median.)

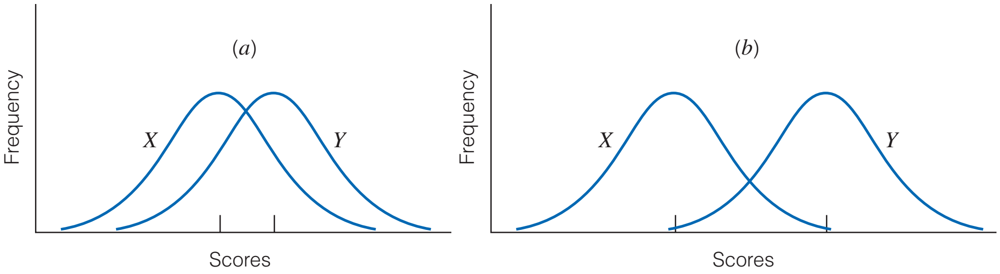


FIGURE 21.1 Distributions that overlap by different amounts.

The test is sometimes referred to as the *Wilcoxon rank-sum test*, in honor of the man who first thought of the basic idea (Wilcoxon, 1945). However, to use the rank-sum test, sample size should be at least 10 for both groups. Mann and Whitney (1947) offered a variant of the same test that we may use when sample sizes are smaller (or greater) than 10 and gave it the name *U*. We will refer to the test by their names to avoid the confusion that might result from the fact that another test described in this chapter is usually known by Wilcoxon's name (Section 21.7).

The Mann-Whitney *U* test is easy to understand at an intuitive level. Suppose there are two groups with distributions as shown in Figure 21.1*a*; we see that they overlap to a substantial extent. Suppose that we combined the two distributions and ranked all scores from 1 (the lowest score) to $n_X + n_Y$ (the highest score). If the distributions overlapped perfectly, and we then summed the ranks of the *X* scores and the ranks of the *Y* scores, we would expect the weighted sum (in the case of unequal *n*) of the ranks to be the same for each group. However, for the distributions in Figure 21.1*a* the ranks of the scores in the *X* group will tend to be somewhat lower than those of the scores in the *Y* group; consequently, the sum of the ranks of the *X* group (**ΣR_X**) will tend to be lower than the sum of the ranks of the *Y* group (**ΣR_Y**). If the two groups are more widely separated, as shown in Figure 21.1*b*, the difference between ΣR_X and ΣR_Y will be even greater. The Mann-Whitney *U* test is based on exactly this difference. It derives from the probability of obtaining a sum of ranks for one distribution that differs from the expected sum of ranks (under the hypothesis of equality of the two distributions) by more than a given amount.

ΣR_X, ΣR_Y symbols for the sum of the ranks of the *X* and *Y* groups

To conduct the test, follow these procedures:

Step 1: Label the two groups *X* and *Y*; *if one group contains fewer cases than the other, it must be labeled* X.

Step 2: Combine all the scores into one distribution of $n_X + n_Y$ cases. Then assign the rank of 1 to the lowest score, 2 to the next lowest, and so on, until all scores are ranked.

Step 3: Find ΣR_X, the sum of the ranks of all scores in the *X* distribution.

Step 4: Calculate U_X as follows:

$$(n_X)(n_Y) + \frac{n_X(n_X+1)}{2} - \sum R_X$$

formula for the Mann-Whitney *U* test

FORMULA FOR THE
MANN-WHITNEY *U*

$$U_X = (n_X)(n_Y) + \frac{n_X(n_X+1)}{2} - \sum R_X \qquad (21.1)$$

When submitting research reports, it is necessary to provide the obtained value of ***U***. However, for convenience, we provide Table J in Appendix D, which gives critical values of ΣR_X

corresponding to one-tailed probabilities of .005, .01, .025, and .05 for n_X values from 3 to 15 and n_Y values from 3 to 15.

Step 5: Locate the row of entries in Table J that corresponds to n_X and n_Y. Then follow procedure (a), (b), or (c), depending on the alternative hypothesis:

- **a.** The alternative hypothesis, H_A, states only that the two population *distributions* are different in *some* way. Find the pair of numbers corresponding to the column heading equal to $\alpha/2$. (If $\alpha = .05$, use the heading designated .025.) Reject H_0 if $\Sigma R_{X\text{calc}}$ equals or falls below the lower number *or* if it equals or exceeds the higher number.
- **b.** H_A: $X < Y$. Find the pair of numbers corresponding to the column heading equal to α. (If $\alpha = .05$, use the heading designated .05.) Reject H_0 if $\Sigma R_{X\text{calc}}$ equals or falls below the lower number.
- **c.** H_A: $X > Y$. Find the pair of numbers corresponding to the column heading equal to α. (If $\alpha = .05$, use the heading designated .05.) Reject H_0 if $\Sigma R_{X\text{calc}}$ equals or exceeds the higher number.

As sample size increases, the distribution of U comes to resemble a normal curve. Thus, when sample sizes are too large to use Table J, we can calculate the Mann-Whitney U test statistic and use the normal curve table (Table A):

$$\frac{U_X - \frac{(n_X n_Y)}{2}}{\sqrt{\frac{n_X n_Y(n_X + n_Y + 1)}{12}}}$$

z formula for the Mann-Whitney *U* test with large samples

MANN-WHITNEY *U* TEST STATISTIC
FOR LARGER SAMPLES

$$z = \frac{U_X - \frac{(n_X n_Y)}{2}}{\sqrt{\frac{n_X n_Y(n_X + n_Y + 1)}{12}}} \tag{21.2}$$

Negative values of z imply $X < Y$, and positive values imply $X > Y$.

Example

Suppose that we give nine children an arithmetic achievement test and record their scores. They then receive daily practice on arithmetic problems. At the end of each practice session, their papers are marked by a teacher's aide, who praises them for the number they got right and states that she thinks they might do even better in the next few days. We treat eight other children the same except for the statements of the teacher's aide. For these children, she points out the problems they missed and tells them that they should be able to do better than that. At the end of 2 weeks, we retest all children on a parallel form of the arithmetic achievement test and calculate their "gain scores" over the 2-week period:

Criticism (X):	−3	−2	0	0	2	5	7	9	
Praise (Y):	0	2	3	4	10	12	14	19	21

Solution

When we examine these scores we see that they are skewed positively. Because the samples are fairly small, we decide to use the Mann-Whitney U test rather than the t test for independent means to test the null hypothesis of no difference between the population distributions. We decide to conduct a two-tailed test at the .05 level of significance.

We must first rank the scores according to Step 2 of the procedures. Note that some scores will be tied in rank. We assign ranks for these scores according to the procedure explained in Section 7.9. We then sum the X ranks:

Criticism (X):	1	2	4	4	6.5	10	11	12	
Praise (Y):	4	6.5	8	9	13	14	15	16	17

$$\sum R_X = 50.5$$

We enter Table J with $n_X = 8$, $n_Y = 9$, and $\alpha = .05$ (for a two-tailed test, we look in the column headed .025). We find that the critical values of ΣR_X are 51 and 93. Because our obtained $\Sigma R_X = 50.5$ falls below the lower critical value, we reject the null hypothesis. The low value of ΣR_X indicates that the gains of the criticized group generally fell below those of the praised group. This is confirmed when we note that the median of the gains of the criticized group is 1, whereas that of the praised group is 10.

If, instead of using Table J, we figure the outcome from Formulas 21.1 and 21.2, we have:

$$U_X = (8)(9) + \frac{8(9)}{2} - 50.5 = 57.5$$

$$z = \frac{57.5 - \frac{(8)(9)}{2}}{\sqrt{\frac{(8)(9)(8 + 9 + 1)}{12}}} = -2.07$$

This value falls just below $z = -1.96$, the critical value of normally distributed z corresponding to $\alpha = .05$ for a two-tailed test.

The Mann-Whitney U test has only two assumptions: (1) independence of samples and (2) continuous scores. With a continuous variable, the possibility of ties is very low. However, the scores recorded in behavioral science research are often discrete rather than continuous, thus creating the possibility of ties. The existence of tied ranks *across* groups reduces the power of the test (i.e., makes it *more* difficult to reject H_0 when it is false). Fortunately, as in the example illustrated, a moderate number of tied ranks does not substantially disturb the outcome.

Glass rank biserial correlation coefficient, (r_g) a measure of effect size for U

What about effect size? The appropriate measure for the U test is the **Glass rank biserial correlation coefficient, r_g**:

FORMULA FOR r_g

$$r_g = \frac{2(\overline{R}_X - \overline{R}_Y)}{n_X + n_Y} \tag{21.3}$$

where: $\overline{R}_X$ = the mean rank for X
$\overline{R}_Y$ = the mean rank for Y.

In our example,

$$r_g = \frac{2(6.31 - 11.39)}{8 + 9}$$

$$= -.60$$

The Glass rank biserial correlation coefficient shows the strength of the relationship between X and Y and can take on values between -1.00 and $+1.00$. Thus, we interpret it much like Pearson's r.

Point of Controversy

A Comparison of the *t* Test and the Mann-Whitney *U* Test with Real-World Distributions

The *t* test was a notable advance in science. It has allowed researchers to apply the precise, objective methods of statistics to small samples of observations that can be gathered quickly and cheaply. Recall, however, that one of the underlying assumptions of the *t* test is that the population from which the sample(s) was drawn is normally distributed. What if the population is not normal in shape?

The traditional answer has been that the *t* test holds up well against violations of the assumption of normality. (Statisticians refer to this as *robustness.*) This conclusion was based largely on work by Boneau (1960), who examined the behavior of the *t* statistic when applied to smooth, continuous nonnormal distributions derived from mathematical models. Although others expressed some serious reservations about use of the *t* test when distributions differed more radically from those studied by Boneau (e.g., Bradley, 1978; Freedman, Pisani, & Purves, 1978, p. A20; Wilkinson, 1988, p. 717), their warnings went unheeded because there was little evidence that real-world distributions differed substantially from normality.

Recall that Micceri (1989), in a study of 440 large, real-world data sets from the social and behavioral sciences, found that none of the distributions were normal in shape, and that very few could even be described as smooth and continuous (see Point of Controversy in Chapter 6). Most distributions of data in the behavioral sciences were multimodal, lumpy, and/or skewed. How well does the *t* test hold up in these real-world situations?

Sawilowsky and Blair (1992) examined the robustness of the independent-samples *t* test with samples drawn from distributions like those found by Micceri. Overall, the *t* test was reasonably robust *to Type I error* when (a) sample sizes were equal (or nearly so), (b) sample sizes were fairly large (25–30), and (c) tests were two-tailed. Differences between nominal and actual alpha levels were usually of a conservative nature. These results agree with those by Boneau (1960). The one notable exception was when distributions were extremely skewed. In this case, the *t* test was nonrobust to Type I error.

But what about robustness to Type II error and power? Earlier studies had found that nonparametric tests were often more robust and provided greater power than the *t* test when samples were drawn from nonnormal populations (Chernoff & Savage, 1958; Neave & Granger, 1968). Sawilowsky and Blair (1992) compared the power for the *t* test and the Mann-Whitney *U* test with small samples (5, 15) drawn from a very asymmetric distribution. The Mann-Whitney *U* test was clearly more powerful.

The *t* test has been in wide use for many decades and researchers will no doubt continue to use it, for traditions end slowly. This text thus devotes considerable attention to the *t* test—but you are now apprised of its limitations. When you must choose a test for comparison of small samples and have doubts about the shape of the population distribution, keep in mind the following conclusion by Sawilowsky and Blair (1992):

> Researchers should not overlook robust nonparametric competitors that are often more powerful than the *t* test when its underlying assumptions are violated.

21.6 An Assumption-Freer Alternative to the *t* Test of a Difference between Two Dependent Groups: The Sign Test

The sign test and the Wilcoxon signed-ranks test are commonly used to test for a difference in central tendency between two *dependent* samples. Both are assumption-freer alternatives to

sign test an assumption-freer test of a difference between two dependent groups (for use when the scale of measurement is ordinal)

the *t* test for dependent means, but each has an important advantage that the other does not. We present the sign test in this section and the Wilcoxon signed-ranks test in the next.

The **sign test** is very useful for analyzing results of studies in which the scale of measurement is ordinal, such as rating degrees of sociability or other aspects of personality. Its name comes from the use of plus or minus signs rather than quantitative measurements (the sign test is *not* a rank-order test).

Example

Suppose that a social psychologist is interested in whether a short course in human relations can enhance an individual's tolerance of people of other ethnic groups. The psychologist worries, however, that intelligence and tolerance of others are correlated variables, so she decides to match 20 pairs of subjects on the basis of their scores on an intelligence test. One member of each pair then takes the course and the other does not (decided randomly). All 40 subjects are then interviewed and rated for their tolerance of others on a 30-point scale. Let us suppose that the results are those shown in Table 21.2 in the columns under "Tolerance score."

Solution

The sign test is appropriate here because the measurement scale is ordinal at best. To use the test, the psychologist assigns a plus sign (+) to a pair of observations when the member who took the course received a higher tolerance score and a minus sign (−) when the opposite was the case, as shown in Table 21.2 in the column labeled "Difference."

If the course made no difference, we would expect there to be as many "pluses" as "minuses," within the limits of sampling variation. In other words, the null hypothesis is that the proportions

TABLE 21.2 *Testing the Difference between Two Dependent Groups: The Sign Test*

	TOLERANCE SCORE		
MATCHED PAIR	*Took course*	*No course*	DIFFERENCE
1	23	21	+
2	17	10	+
3	15	17	−
4	30	24	+
5	25	18	+
6	18	12	+
7	10	9	+
8	24	20	+
9	16	17	−
10	30	27	+
11	15	9	+
12	20	18	+
13	25	20	+
14	22	17	+
15	20	21	−
16	16	12	+
17	20	17	+
18	28	24	+
19	30	28	+
20	24	21	+

of pluses and minuses in the population are equal. We can apply chi-square to this type of question. The expected and observed frequencies of plus and minus are as follows:

	(−)	(+)
f_o	3	17
f_e	10	10

We calculate chi-square as described in Section 20.2:

$$\chi^2 = \frac{(f_{o+} - f_{e+})^2}{f_{e+}} + \frac{(f_{o-} - f_{e-})^2}{f_{e-}}$$

$$= \frac{(17 - 10)^2}{10} + \frac{(3 - 10)^2}{10}$$

$$= 9.80$$

With 1 degree of freedom, χ^2 must reach or exceed 3.84 to be significant at the 5% level and 6.64 to be significant at the 1% level. (These values are determined from Table I in Appendix D for a nondirectional test.) The psychologist's χ^2 surpasses both of these critical values, so it appears that the pluses predominate because of some systematic process—perhaps the human-relations course resulted in increased understanding of others.

If you use the sign test, a difference will occasionally be zero and cannot therefore be categorized as a plus or a minus. We may solve this dilemma in one of several ways. Probably the simplest is to ignore such cases, reduce *n* accordingly, and proceed with the test on the remaining values.

Evaluation of the sign test by χ^2 will give reasonable accuracy for 10 or more pairs of scores. *With fewer than 10 pairs, we should use the binomial test in place of chi-square.*

An assumption required for the sign test is that no difference is exactly zero. The use of the method as described here will be reasonably satisfactory provided the number of zeros is small.

The special advantages of the sign test are its ease of application and assumptions that are less restrictive than those for the Wilcoxon signed-ranks test. Its main disadvantage is that it is less sensitive because it disregards some information that the other takes into account. In our example, each plus simply indicates that the subject who completed the human-relations course is more tolerant than the paired subject who did not take the course, but *no account is taken of how much more.*

21.7 Another Assumption-Freer Alternative to the *t* Test of a Difference between Two Dependent Groups: The Wilcoxon Signed-Ranks Test

Wilcoxon signed-ranks test
a rank-order test of a difference between two dependent groups

The **Wilcoxon signed-ranks test** is an alternative to the sign test (and also to the *t* test of the difference between two dependent means). It is more powerful than the sign test because it weighs the size of a difference between pairs of scores. However, it demands an assumption that we may not be willing to make (more about this later).

Suppose we are studying the effect of a drug on the performance of mental operation—for example, on the speed and accuracy of the addition and subtraction of figures. We form 10 pairs of subjects, matched according to their pretest score earned on similar arithmetic problems, and then randomly assign members of each pair to two treatments: the drug

TABLE 21.3 *Testing the Difference between Two Dependent Samples: The Wilcoxon Signed-Ranks Test*

	①	①	②	③	④	⑤	
PAIR	**PLACEBO (X)**	**DRUG (Y)**	**$\lvert X - Y \rvert$**	**RANK OF $\lvert X - Y \rvert$**	**SIGNED RANK OF $X - Y$**	**R_+**	**R_-**
1	24	28	4	2	-2		2
2	39	29	10	6.5	$+6.5$	6.5	
3	29	34	5	3.5	-3.5		3.5
4	28	21	7	5	$+5$	5	
5	25	28	3	1	-1		1
6	32	15	17	10	$+10$	10	
7	31	17	14	8	$+8$	8	
8	33	28	5	3.5	$+3.5$	3.5	
9	31	16	15	9	$+9$	9	
10	22	12	10	6.5	$+6.5$	6.5	
				$\sum \lvert X - Y \rvert = 55$		$R_+ = 48.5$	$R_- = 6.5$
				$= n(n + 1)/2$		$T = 6.5$	

and a placebo. At an appropriate time after ingesting the substance (active or inert), the subjects take the test. Their scores are recorded in the columns labeled "Placebo (X)" and "Drug (Y)" in Table 21.3. We wish to know whether the drug made a difference in performance; we decide to conduct a two-tailed test at $\alpha = .05$ and to use the Wilcoxon signed-ranks test. We describe the procedure below and show the calculations in Table 21.3.

Step 1: Record the paired scores in two columns. In the table, these appear under the column headings "Placebo (X)" and "Drug (Y)." See ①.

Step 2: Obtain the difference between members of each pair ($X - Y$). See ②.

Step 3: If any difference is zero, disregard it in subsequent calculations and reduce n, the number of pairs of scores, accordingly.

Step 4: *Disregard the sign of the difference obtained*, and then supply ranks to the absolute magnitude of the differences, assigning a rank of 1 to the smallest of the differences, 2 to the next smallest, and so forth. Ties in rank are handled as described in Section 7.9. The outcome is shown in the column headed "Rank of $\lvert X - Y \rvert$." See ③.

T
the sum of the ranks with a negative sign (**R_-**) or the sum of the ranks with a positive sign (**R_+**), whichever is smaller

Step 5: Next, resupply the appropriate sign of the differences to the rank of the differences. We show this in the column labeled "Signed Rank of $X - Y$." See ④.

Step 6: The test statistic, ***T***, will be the sum of the ranks with a negative sign or the sum of the ranks with a positive sign, *whichever is smaller.* We will call the former **R_-** and the latter **R_+**. Often it will be obvious which to calculate; for purposes of completeness, we show both in Table 21.3. See ⑤.

Step 7: In Table K in Appendix D, you will find critical values for one- and two-tailed tests at the usual levels of significance. We compare the obtained value of T (*disregarding the sign*) with the entry in the table corresponding to n (the number of pairs), α, and the nature of our alternative hypothesis. For a two-tailed test, we reject the null hypothesis if the calculated value *equals or is less than the tabulated value.* For a one-tailed test, we must also take the sign of the test statistic into account. A positive sign means that X is greater than Y; a negative sign means the opposite. However, because our test statistic is the smaller of the two sums, rejecting the null hypothesis for a very small sum of negative ranks means that positive ranks predominate and that therefore $X > Y$. Similarly, if the test statistic is positive, it indicates that $X < Y$.

In our example, R_- is smaller than R_+ and is therefore the test statistic. According to Table K (in Appendix D), the critical value for a two-tailed test at $\alpha = .05$ with 10 pairs of scores is 8. Because the absolute value of R_- is smaller, we reject the null hypothesis and conclude that the drug resulted in reduced performance.

When the sample size is too large to use Table K, the statistic z is approximately normally distributed:

$$\frac{\boldsymbol{R_+ - .25n(n+1)}}{\sqrt{\dfrac{\boldsymbol{n(n+1)(2n+1)}}{\mathbf{24}}}}$$

z formula for the Wilcoxon signed-ranks test with large samples

z STATISTIC FOR THE WILCOXON SIGNED-RANKS TEST

$$z = \frac{R_+ - .25n(n+1)}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \tag{21.4}$$

If it is more convenient, substitute R_- (again, disregarding its sign) for R_+ in the formula.

You can see why the Wilcoxon signed-ranks test is more sensitive than the sign test. The sign test responds only to the direction of the difference between a pair of scores, whereas the Wilcoxon test uses additional information about the size of the difference. For useful interpretation of the outcome of the Wilcoxon test, we must assume not only that the scores are from an ordinal scale of measurement, but also that the *differences between pairs of scores can be placed in rank order.* In behavioral science, this may often be a hazardous assumption. Is it true that the difference between scores of 20 and 25 is the same as that between scores of 40 and 45? This depends on the scale properties of the measure, which are often unknown. See Section 1.6 for further comment on this matter.

Other assumptions for the test are the random assignment of treatment condition to members of a pair (with independent assignment among different pairs), no differences of zero, and no ties in rank.

matched-pairs rank biserial correlation coefficient, (r_C) a measure of effect size for the Wilcoxon signed-ranks test

To calculate effect size, we use the **matched-pairs rank biserial correlation coefficient, r_C**:

FORMULA FOR r_C

$$r_C = \frac{4\left|T - \left(\dfrac{R_+ + R_-}{2}\right)\right|}{n(n+1)} \tag{21.5}$$

In our example,

$$\begin{aligned} r_C &= \frac{4\left|6.5 - \left(\dfrac{48.5 + 6.5}{2}\right)\right|}{10(11)} \\ &= .76 \end{aligned}$$

We calculate the magnitude of r_C the same way we do for r.

21.8 An Assumption-Freer Alternative to the One-Way ANOVA for Independent Groups: The Kruskal–Wallis Test

Kruskal–Wallis test
a rank-order test of a difference among three or more independent groups

Look again at the assumptions associated with the one-way ANOVA (Section 18.11). The **Kruskal–Wallis test** is an alternative to ANOVA (with independent samples) for situations in which measurements are on an ordinal scale or when we have doubts about the distributional assumptions associated with the F test. You may think of it as an extension of the Mann-Whitney U test to more than two groups. As with the Mann-Whitney test, the null hypothesis concerns the identity of the population distributions rather than the identity of some particular measure of central tendency for the populations. Under ordinary circumstances, however, it is a good test for central tendency. (Again, the median is probably the best statistic to use when examining a significant outcome.)

Suppose we combine the scores from k groups and rank all scores from 1 (the lowest score) to $n_1 + n_2 + \cdots + n_k$ (the highest score). Look at Section 7.9 if you need to review how to rank scores. After separating the groups again, we find the mean rank of each and then compare them. The basic logic involves determining whether the mean ranks ($\Sigma R/n$) for the groups differ more than would be expected due to random sampling variation alone. If the samples all came from identical populations, we would expect the mean ranks of the groups to vary to some extent (due to random variation) around the overall rank mean. If, however, the samples come from different populations, we would expect the mean ranks of the groups to vary more widely. The Kruskal–Wallis test examines the magnitudes of the k differences between mean rank of each group and the overall mean rank and finally compares the total magnitude of the differences with what might be expected by chance. Notice the similarity of the rationale for the one-way ANOVA and the rationale here. (Figures 18.1 and 18.2 are worth studying again because they present pictorially the essence of the argument developed here.)

H
the test statistic for the Kruskal–Wallis test

The test statistic, ***H***, for the Kruskal–Wallis test is:

THE KRUSKAL–WALLIS TEST STATISTIC, H

$$H = -3(n_{\text{total}} + 1) + \frac{12}{n_{\text{total}}\,(n_{\text{total}} + 1)}\left[\frac{\left(\sum R_1\right)^2}{n_1} + \frac{\left(\sum R_2\right)^2}{n_2} + \cdots + \frac{\left(\sum R_k\right)^2}{n_k}\right] \qquad (21.6)$$

ΣR_1
sum of the ranks in group 1

where: ΣR_1 = the sum of the ranks in Group 1, and so forth
n_1 = the number of cases in Group 1, and so forth
n_{total} = the number of cases in all the groups combined.

When H_0 is true, the sampling distribution of H approximates the chi-square distribution (Chapter 20). With three groups and five or more cases per group, we may use the chi-square distribution to evaluate H with good approximate results. (With more than three groups, some groups can have as few as two or three cases.) We compare our H_{calc} with the critical values of χ^2 in Table I in Appendix D for $df = k - 1$, where k is the number of groups. Because H reflects the magnitude of the differences and is not sensitive to their direction, the region of rejection lies entirely in the upper tail of the χ^2 distribution. If we consult the column of Table I headed by .05, for example, it supplies the critical value of χ^2 (or H) for a nondirectional test at $\alpha = .05$. We reject the null hypothesis for H_{calc} that equals or exceeds the tabulated value.

One assumption of the Kruskal–Wallis test is that there are no ties in rank. As with the Mann-Whitney U test for two independent groups, the effect of ties in rank is not great unless there are many of them.

Example

Suppose we study three brands of computer printers, A, B, and C, for their legibility. We sample six brand A printers, six brand B, and five brand C. The legibility scores of the 17 printers sampled are shown in Table 21.4.

Solution

Using the Kruskal–Wallis test to examine possible differences, the first step (①) is to combine the 6 + 6 + 5 = 17 scores into a single distribution and to assign the scores ranks beginning with 1 for the lowest score and ending with 17 for the highest. Notice that the score of 26 appears three times, so we assign the mean of the three ranks available for these scores (8) to each (see Section 7.9). We calculate ΣR, the sum of the ranks for each group, and record it below the body of the table (Step ②). We check the assignment of ranks (Step ③) using the fact that:

$$\sum^{\text{all}} R = \frac{n_{\text{total}}(n_{\text{total}} + 1)}{2}$$

TABLE 21.4 *Examination of the Difference in Legibility of Three Brands of Computer Printers Using the Kruskal-Wallis Test*

PRINTER A ①		PRINTER B ①		PRINTER C ①	
Score	*Rank*	*Score*	*Rank*	*Score*	*Rank*
22	3	26	8	23	4
24	5	36	16	30	12
21	2	27	10	29	11
17	1	37	17	26	8
26	8	33	15	31	13
25	6	32	14		
$n_1 = 6$	$\sum R_1 = 25$	$n_2 = 6$	$\sum R_2 = 80$	$n_3 = 5$	$\sum R_3 = 48$

Calculations:

② $\sum^{\text{all}} R = 25 + 80 + 48 = 153$

③ $\sum^{\text{all}} R = \dfrac{n_{\text{total}}(n_{\text{total}} + 1)}{2} = \dfrac{17(17 + 1)}{2} = 153$ (check)

④
$$H = -3(n_{\text{total}} + 1) + \frac{12}{n_{\text{total}}(n_{\text{total}} + 1)}\left[\frac{\left(\sum R_1\right)^2}{n_1} + \frac{\left(\sum R_2\right)^2}{n_2} + \frac{\left(\sum R_k\right)^2}{n_k}\right]$$

$$= -3(17 + 1) + \frac{12}{17(17 + 1)}\left[\frac{(25)^2}{6} + \frac{(80)^2}{6} + \frac{(48)^2}{5}\right]$$

$$= -54 + .0392(1631.63)$$

$$= 9.96$$

⑤ $df = k - 1 = 3 - 1 = 2$

⑥ $\chi^2_{\text{crit}} = 5.99$ for $\alpha = .05$

Decision:
⑦ H_0 is rejected at $\alpha = .05$.

This is shown immediately below the sum of the ranks for the groups. Next we calculate H, which yields a value of 9.96 (Step ④). Using the χ^2 distribution to evaluate H, $df = k - 1 = 3 - 1 = 2$ (Step ⑤). For $\alpha = .05$ and $df = 2$, Table I in Appendix D gives $\chi^2_{\text{crit}} = 5.99$ (Step ⑥). Our $H_{\text{calc}} = 9.96$ falls beyond that value, so we reject the null hypothesis (Step ⑦).

We are not limited to the overall test of significance; as with analysis of variance, we can make multiple comparisons if our H proves to be significant (Dunn, 1964). To make comparisons of all possible pairs of mean ranks, use the Mann–Whitney U test. However, we must use a revised alpha level to account for the increased chance of a Type I error. Dunn suggests that we use *alpha/number of pairs of mean ranks tested*. In our example there are three possible comparisons (A versus B, A versus C, B versus C), so our revised alpha level is .05/3 = .017.

epsilon-squared (E^2_R) a measure of effect size for the Kruskal–Wallis test

An appropriate measure of effect size for the Kruskal–Wallis test is **epsilon-squared, E^2_R**:

FORMULA FOR EPSILON-SQUARED

$$E^2_R = \frac{H}{(n^2_{\text{total}} - 1)/(n_{\text{total}} + 1)} \qquad (21.7)$$

Epsilon-squared can have values between 0 (no relationship) and 1.00 (indicating a perfect relationship between the independent and dependent variables). In our example,

$$\begin{aligned} E^2_R &= \frac{9.96}{(17^2 - 1)/(17 + 1)} \\ &= \frac{9.96}{288/18} \\ &= .623 \end{aligned}$$

21.9 An Assumption-Freer Alternative to ANOVA for Repeated Measures: Friedman's Rank Test for Correlated Samples

Friedman's rank test a rank-order test of a difference among three or more dependent groups

χ^2_F statistic for Friedman's rank test

$\frac{12\sum R^2_j}{Nk(k+1)} - 3N(k+1)$ formula for Friedman's rank test for correlated samples

If you want to use a repeated-measures design, and you have serious doubts about the distributions in the populations from which you have drawn your samples, an assumption-freer alternative to ANOVA is **Friedman's rank test (χ^2_F)** for correlated samples. Friedman's test (named after economist Milton Friedman, who developed it) may be used when there are at least three levels in the repeated-measures design and when scores are in the form of ranks:

$$\chi^2_F = \frac{12\sum R^2_j}{Nk(k+1)} - 3N(k+1) \qquad (21.8)$$

where: χ^2_F = a statistic with a sampling distribution that closely resembles the chi-square distribution with $k - 1$ degrees of freedom
R^2_j = the sum of the ranks in the jth condition
N = sample size
k = number of conditions

TABLE 21.5 *Data, Ranks, and Initial Calculations for Friedman's Test*

SUBJECT	NO STRESS	MILD STRESS	MODERATE STRESS
1	65 (3)	60 (2)	50 (1)
2	80 (2)	85 (3)	70 (1)
3	75 (3)	60 (1)	65 (2)
4	40 (3)	35 (2)	20 (1)
5	60 (2)	65 (3)	50 (1)
6	70 (1)	75 (2)	80 (3)
7	90 (2)	95 (3)	70 (1)
8	55 (3)	40 (2)	25 (1)
9	85 (2)	90 (3)	75 (1)
10	85 (1)	90 (2)	95 (3)
11	35 (3)	30 (2)	15 (1)
12	75 (3)	65 (2)	60 (1)
	$R_1 = 28$	$R_2 = 27$	$R_3 = 17$

To use Friedman's test, you must first rank the scores *for each subject across all conditions.* When there are ties in scores, assign the mean rank to both. As an example, suppose we are interested in knowing the effects of stress on the ability to solve problems. Twelve subjects are recruited and asked to solve problems under each of three conditions—after a pretest period of no stress, mild stress, or moderate stress. We counterbalance the three conditions to ensure that the order of presentation does not affect our results, and the problems are graded on a 100-point scale. Hypothetical results are presented in Table 21.5, accompanied by their ranks. Look, for example, at subject 1. His or her scores of 65 after no stress, 60 after mild stress, and 50 after moderate stress correspond to ranks of 3, 2, and 1, respectively.

The null hypothesis is that the scores under each condition are from identical populations. If this were true, the rankings within each condition would be random, and any differences across conditions would be due to chance. Let us now solve for χ_F^2:

$$\begin{aligned}\chi_F^2 &= \frac{12(28^2 + 27^2 + 17^2)}{12(3)(4)} - 3(12)(4) \\ &= \frac{12(1802)}{144} - 144 \\ &= \frac{21{,}624}{144} - 144 \\ &= 150.17 - 144 \\ &= 6.17\end{aligned}$$

To evaluate the outcome at $\alpha = .05$, look in Table I of Appendix D for $df = k - 1 = 2$. The critical value is 5.99. Our calculated value of 6.17 exceeds this, so we reject the null hypothesis and conclude that ability to solve problems is affected by stress. As with ANOVA, once χ_F^2 is found to be significant, we may make multiple comparisons (Dunn, 1964). We use the Wilcoxon signed-ranks test, but to account for the increased possibility of Type I errors, the alpha level should be revised by *alpha/number of pairs rank sum tested.* In our example, alpha is .05/3 = .017.

concordance coefficient (W) a measure of effect size for Friedman's rank test

To measure effect size, calculate the **concordance coefficient, W**:

FORMULA FOR THE CONCORDANCE COEFFICIENT

$$W = \frac{\chi_F^2}{N(k - 1)} \tag{21.9}$$

In our example,

$$W = \frac{6.17}{12(2)} = .257$$

The concordance coefficient takes on values between 0 and 1.00, and in this case indicates a moderate effect.

21.10 Summary

To use t, F, or χ^2, we must make assumptions about the underlying population distributions. The two most important assumptions are that the populations are normally distributed and that there is homogeneity of variance. With moderate to large samples, these "parametric" tests are reasonably accurate (because of the central limit theorem) even when the assumptions are violated. The major problem comes with the use of small samples, for which violation of the assumptions can lead to considerable inaccuracy.

Assumption-freer tests make no assumption about the underlying distribution and are more powerful than parametric tests when sample size is small and the population distribution is not normal. Most of the assumption-freer tests introduced in this chapter are rank-order tests in which original scores are converted to ranks. The rank-order tests were derived from randomization tests that compare the obtained outcome with the distribution of outcomes from all possible combination of scores.

The Mann-Whitney U test is commonly used in place of the t test for independent samples. With more than two groups, we can use the Kruskal–Wallis test in place of one-way ANOVA. Although the null hypothesis for these tests is that the population distributions are the same, we can use them for detecting differences in central tendency if we are relatively certain that differences in variability and shape are not pronounced.

The sign test and Wilcoxon's signed-ranks test are popular alternatives to the t test for dependent samples. Wilcoxon's test is the more powerful of the two, but it assumes that the differences between pairs of scores can be placed in rank order. With more than two groups, we can use Friedman's rank test in place of repeated-measures ANOVA.

Key Terms, Concepts, and Symbols

nonparametric statistics (p. 382)
distribution-free methods (p. 382)
assumption-freer tests (p. 382)
randomization tests (p. 383)
rank-order tests (p. 385)
bootstrap method (p. 386)
Mann-Whitney U test (p. 386)
ΣR_X, ΣR_Y (p. 387)
U (p. 387)
Glass rank biserial correlation coefficient (p. 389)
r_g (p. 389)
sign test (p. 391)
Wilcoxon signed-ranks test (p. 392)
T (p. 393)
R_+, R_- (p. 393)
matched-pairs rank biserial correlation coefficient r_C (p. 394)

Kruskal–Wallis test (p. 395)
H (p. 395)
$\Sigma R_1, \Sigma R_2$ (p. 395)
epsilon-squared (p. 397)
E^2_R (p. 397)
Friedman's rank test (p. 397)
χ^2_F (p. 397)
concordance coefficient *W* (p. 399)

Problems

Note: Answers to odd-numbered questions are in Appendix C. Answers to even-numbered questions are in the Instructor's Manual.

1. Explain the differences between the terms *parametric test, nonparametric test, distribution-free test,* and *assumption-freer test.*

2. (a) Why might a researcher choose to use assumption-freer methods? (b) Would sample size influence his or her decision? (c) Are the assumption-freer methods truly assumption free? Explain.

3. How does the sampling distribution of a randomization test differ from the sampling distributions of *t*, *F*, and χ^2?

4. What is the difference between the null hypothesis tested by parametric tests (e.g., *t*) and the null hypothesis tested by assumption-freer tests?

5. With regard to outliers, explain the advantage of using a rank-order test.

6. Place the following scores in rank order: 97, 105, 110, 92, 90, 100, 105, 135, 102, 96. If you need to review, look at Section 7.9.

7. In a large class, 13 students required a make-up examination. The instructor gave the same test to everyone but because of circumstances, he could find a quiet environment for only 5 of the students. The other 8 had to take their test in a noisy environment. Assignment to testing environment can be considered random. The scores of the two groups were

Noisy environment:	55	38	46	63	35	61	56	47
Quiet environment:	62	56	47	65	59			

Is the difference in performance significant according to a two-tailed test using $\alpha = .05$? Use the Mann-Whitney *U* test to solve the problem. Interpret the outcome using appropriate descriptive statistics.

8. Rework Problem 7 using the large sample (*z*) approach.

9. *Androgyny* sometimes refers to the ability of an individual to fill both "masculine" and "feminine" gender roles. An androgynous male, for example, would have no difficulty helping his female partner clean house or do the cooking. A sex-stereotyped male, on the other hand, might refuse to engage in these activities. Suppose that a marriage counselor believes that androgyny is a characteristic that helps lead to long-lasting relationships. To test this, she administers a test to measure androgyny (higher scores indicate greater androgyny) to men whose marriages broke up within 3 years and to men whose marriages have lasted at least 10 years. Suppose the scores were as follows:

Less than 3:	4	6	2	5	8	3	1	
At least 10:	9	11	7	6	14	13	12	10

Apply the Mann-Whitney *U* test to evaluate these results. What can the counselor conclude?

10. Suppose we had doubts about the normality of the two distributions of scores in Problem 7 in Chapter 14. Rework the problem using the Mann-Whitney *U* test with $\alpha = .05$. Did you get the same outcome?

11. In two equivalent problem situations, 12 children are praised on one occasion and blamed on another. Their reactions toward others subsequent to the problem situations are judged for aggressiveness. Ratings of aggressive behavior are as follows:

Child:	A	B	C	D	E	F	G	H	I	J	K	L
Praise:	6	8	2	1	4	5	6	9	8	7	7	3
Blame:	7	11	4	5	3	7	9	8	9	11	9	6

Examine, by the sign test, the hypothesis of no difference between the two treatments. Use $\alpha = .05$ and $\alpha = .01$.

12. Eleven pipe smokers were blindfolded and given a pipe of tobacco to smoke. At a later session, the same subjects were again blindfolded and given a pipe to smoke. This time the pipe contained a heated coil rather than tobacco, so the subjects drew in nothing but warm air. At both sessions, blood pressure was measured before and after puffing the pipe. Below is a record of change in blood pressure during each session (a positive number represents an increase, a negative number a decrease):

Subject:	A	B	C	D	E	F	G	H	I	J	K
First session (tobacco):	12	4	6	17	14	4	20	7	0	10	7
Second session (warm air):	14	4	6	14	6	5	11	2	−4	3	7

(a) Test for a difference between scores at the first session and those at the second. Use the sign test to make a two-tailed test at the 5% significance level. (b) What appears to be interesting about the data aside from the result of the comparison you made in (a)? (c) What comments do you have about the way the study was conducted?

13. Eleven male subjects responded to an attitude questionnaire designed to evaluate their tendency to view women in stereotypic roles. A month later they were shown a film that illustrated the ways in which women have often been treated unfairly relative to men. Four months later the subjects again responded to the attitude questionnaire. The subjects' scores on the questionnaire (the higher the score, the greater the bias) before and after were as follows:

Subject:	A	B	C	D	E	F	G	H	I	J	K
Before:	21	25	31	26	21	31	28	24	27	24	22
After:	13	26	35	24	17	24	22	27	17	19	18

(a) Conduct a nondirectional test for a difference in response between the two sets of data. Use the Wilcoxon signed-ranks test and the 5% significance level. (b) Calculate effect size. (c) Do you feel that the way the study was conducted was adequate? Explain.

14. Rework Problem 13a using the large-sample (z) Wilcoxon approach.

15. (a) Repeat Problem 12a, but evaluate the outcome with the Wilcoxon signed-ranks test. (b) Calculate effect size.

16. Fifteen applicants to the Better Ways Corporation were randomly assigned to one of three new training programs. After completing the programs, all were placed in the same type of job. Six weeks later a supervisor rated each of them on their job performance. Suppose the ratings were as shown in Data 21A. Apply the appropriate nonparametric test to evaluate these results. What can the Better Ways Corporation conclude?

DATA 21A

GROUP 1	GROUP 2	GROUP 3
2	20	10
4	18	17
3	14	16
1	8	12
6	15	13

17. You are interested in knowing whether morning eating habits affect energy levels later in the day. Nine subjects are tested under three conditions in a counterbalanced design: no breakfast, coffee only, or juice plus cereal. The following are hypothetical self-ratings taken at 2 p.m. The scale ranges from 1 (very little energy) to 5 (very energetic). Use Friedman's test to evaluate the results and compute effect size.

DATA 21B

SUBJECT	NO BREAKFAST	COFFEE	JUICE + CEREAL
1	1	2	4
2	2	1	3
3	3	3	5
4	4	2	3
5	1	3	2
6	2	1	5
7	3	5	4
8	1	2	4
9	2	3	5

This page intentionally left blank

EPILOGUE

The Realm of Statistics

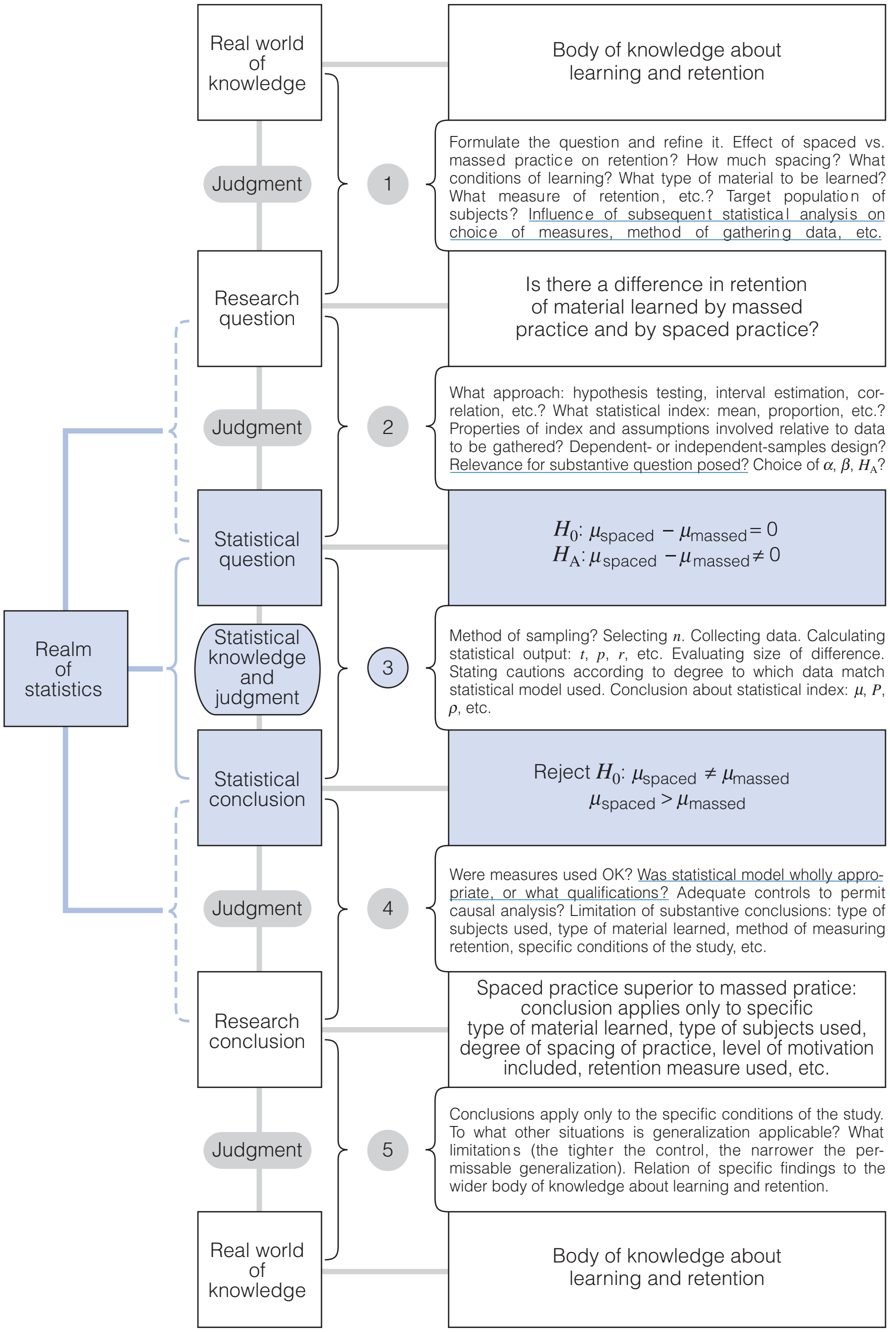
research question a question of fact about the subject matter under investigation

statistical question a question about a numerical aspect of the observations

We developed an outline of the role of applied statistics early in this book. In Chapter 1 we considered the difference between a research question (for example, Do children learn better with spaced practice or massed practice?) and a statistical question (Does $\mu_{\text{spaced}} = \mu_{\text{massed}}$?) and the difference between a statistical conclusion and a research conclusion. We also explained how statistical procedures are only an intermediate part of an investigation. Now that you have learned so much more about the nature of statistical procedures, it is time to return to that overall picture and examine it at a more sophisticated level than was possible earlier.

Figure E.1 shows the typical steps in the development of a study in the behavioral sciences. The square boxes on the left show the progression in general terms, from the real world of knowledge at the beginning to the real world of knowledge again at the end. To the right of each of these boxes (connected by a straight line) is an example of that phase for a hypothetical study of learning and retention.

Between the square boxes are representative descriptions of the kinds of actions we must take to progress from one box to the next. We list these by numbers: ① through ⑤. Notice the word *representative*; it is not possible to list all the considerations that investigators must take into account. Many of the statements or questions we identify here involve not just a single step but also the interaction between factors characterizing one step and those pertaining to another. In some cases, we have underlined statements or questions. These identify considerations that particularly demonstrate the interplay of statistical and research questions.



Body of knowledge about learning and retention

Formulate the question and refine it. Effect of spaced vs. massed practice on retention? How much spacing? What conditions of learning? What type of material to be learned? What measure of retention, etc.? Target population of subjects? Influence of subsequent statistical analysis on choice of measures, method of gathering data, etc.

Is there a difference in retention of material learned by massed practice and by spaced practice?

What approach: hypothesis testing, interval estimation, correlation, etc.? What statistical index: mean, proportion, etc.? Properties of index and assumptions involved relative to data to be gathered? Dependent- or independent-samples design? Relevance for substantive question posed? Choice of α, β, H_A?

H_0: $\mu_{\text{spaced}} - \mu_{\text{massed}} = 0$
H_A: $\mu_{\text{spaced}} - \mu_{\text{massed}} \neq 0$

Method of sampling? Selecting n. Collecting data. Calculating statistical output: t, p, r, etc. Evaluating size of difference. Stating cautions according to degree to which data match statistical model used. Conclusion about statistical index: μ, P, ρ, etc.

Reject H_0: $\mu_{\text{spaced}} \neq \mu_{\text{massed}}$
$\mu_{\text{spaced}} > \mu_{\text{massed}}$

Were measures used OK? Was statistical model wholly appropriate, or what qualifications? Adequate controls to permit causal analysis? Limitation of substantive conclusions: type of subjects used, type of material learned, method of measuring retention, specific conditions of the study, etc.

Spaced practice superior to massed pratice: conclusion applies only to specific type of material learned, type of subjects used, degree of spacing of practice, level of motivation included, retention measure used, etc.

Conclusions apply only to the specific conditions of the study. To what other situations is generalization applicable? What limitations (the tighter the control, the narrower the permissable generalization). Relation of specific findings to the wider body of knowledge about learning and retention.

Body of knowledge about learning and retention

FIGURE E.1 The steps of a study in the behavioral sciences.

Next, notice that between the square boxes there is a reminder that some form of judgment is required at each step. Statistical calculations, once one has learned how to do them, follow a "formula" that can be laid out in a simple sequence of cook-book instructions. But at its very best, this gets us only from the statistical question to the statistical conclusion. In all the other steps in the progression (and to some extent in the one just mentioned), judgment based on knowledge of the research discipline enters in most importantly. This reminds us that statistical procedures do not carry automatic authority but serve as merely one step in the chain.

> Distinguishing statistical significance from theoretical significance (Kirk, 1996) will help the entire research community publish more substantial results. . . . Good theories and intelligent interpretation advance a discipline more than rigid methodological orthodoxy. . . . Statistical methods should guide and discipline our thinking but should not determine it. (Wilkinson and the Task Force on Statistical Inference, 1999)

statistical conclusion a conclusion about a numerical property of the data

We have indicated the realm of statistics at the extreme left of the figure. Its central domain (shown in purple) begins with the formulation of the statistical question and ends with the statistical conclusion. Nevertheless, we should take statistical considerations into account when moving from the research question to the statistical question and again when moving from the statistical conclusion to the research conclusion.

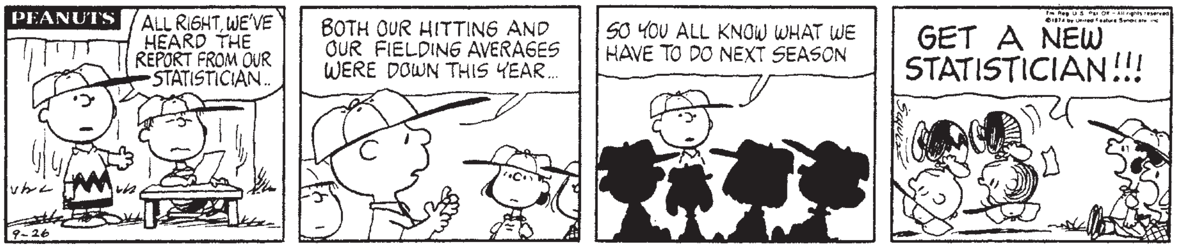
research conclusion a conclusion about the subject matter

The move from the statistical conclusion to the research conclusion is a particularly important matter. Suppose, for example, that we wish to compare the performance of two groups. In the statistical conclusion of the study, we may be informed that mean performance is different for the two groups. But in moving to the research conclusion, we want to know *why* that difference occurred. To what do we attribute causality? The fact that a difference has occurred does not inform us as to the reason for it. It will not be clear that the difference in experimental conditions we imposed on the two groups is responsible for the difference *unless adequate controls* were used. The discussion about controls in Section 14.13 and those concerning causation in Sections 7.10 and 8.10 are particularly important here.

After we have made a conclusion about causality, we must then concern ourselves with generalizations (Sections 13.12 and 14.12). As indicated by ⑤ in Figure E.1, research conclusions apply, strictly speaking, only to the *particular* circumstances that characterize the study. See ① in the figure for examples of such specification. The problem in generalizing to other types of subjects, kinds of material learned, and conditions of learning is serious because (again, strictly speaking) it is necessary to demonstrate that any differences between the characteristics of the situation in which the study was conducted and those of the situation to which we wish to generalize would have no effect—a tall order indeed. This is where great experience with the variables concerned can be very important in suggesting what can be safely ventured and what cannot.

The general (and ideal) model for the two-group (or more) comparisons developed in this book is founded on the concept of drawing two (or more) random samples from a fully defined population and treating one group one way and the other another. In practice, we are seldom so blessed but must instead take a convenient group at hand and randomly assign these subjects to two (or more) treatment groups. Now, *both* random sampling and random assignment serve the purpose of experimental control, as we described in Section 14.13. Whereas we may properly generalize the result of random sampling to the defined target population, the outcome of a study using random assignment applies *only* to the kind of subjects represented in the study. This is not to say that we cannot make generalizations, only that statistics does not provide a basis for making those generalizations.

Finally, in Figure E.2 we provide a flowchart that summarizes the various considerations leading to the choice of a particular statistical procedure. We hope this helps should you ever engage in a research project yourself.


PEANUTS
ALL RIGHT, WE'VE HEARD THE REPORT FROM OUR STATISTICIAN...
9-26
BOTH OUR HITTING AND OUR FIELDING AVERAGES WERE DOWN THIS YEAR...
SO YOU ALL KNOW WHAT WE HAVE TO DO NEXT SEASON
GET A NEW STATISTICIAN!!!

PEANUTS: © United Feature Syndicate, Inc.

LibraryPirate

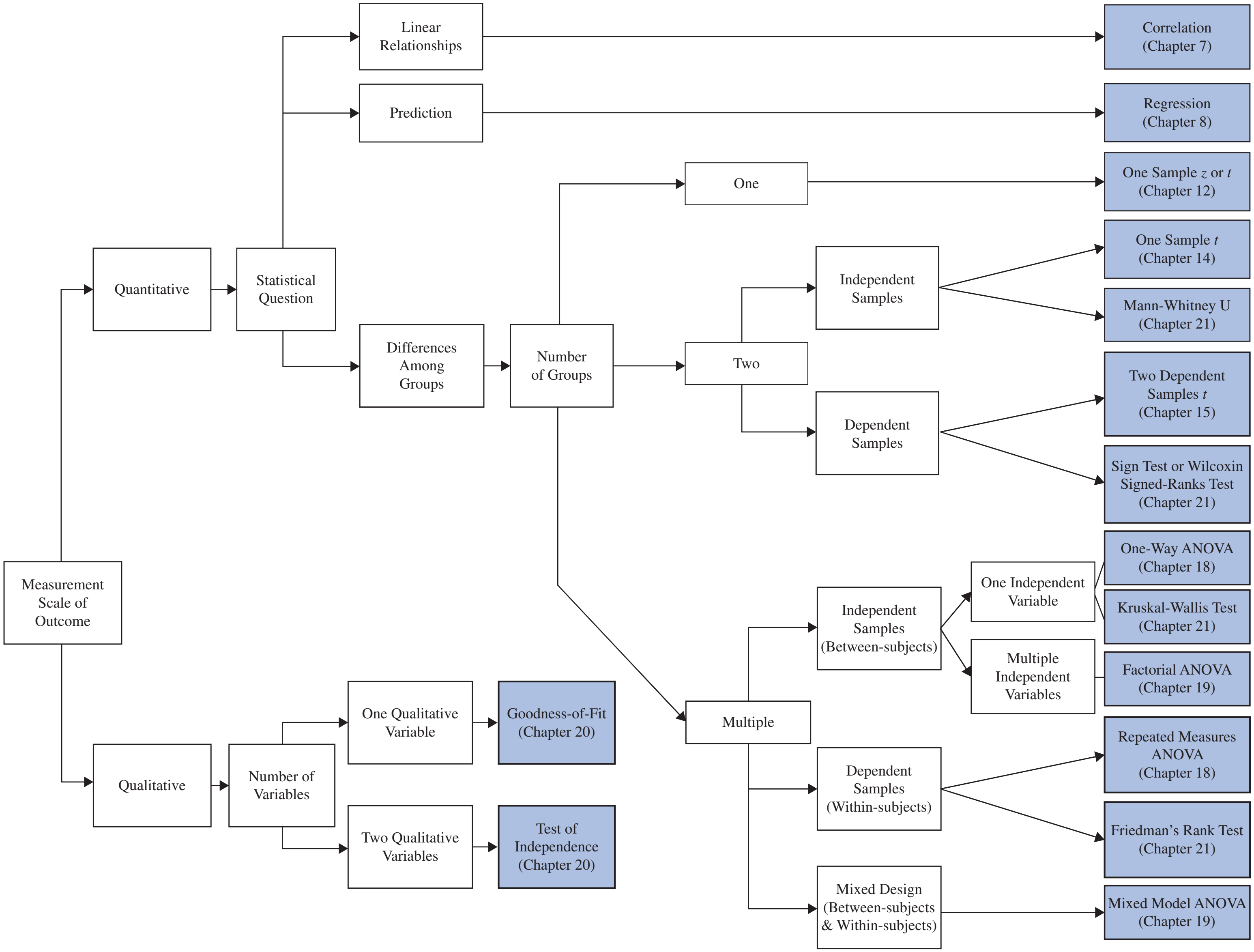

Measurement Scale of Outcome
Quantitative
Statistical Question
Linear Relationships
Prediction
Differences Among Groups
Number of Groups
One
Two
Multiple
Independent Samples
Dependent Samples
Independent Samples (Between-subjects)
Dependent Samples (Within-subjects)
Mixed Design (Between-subjects & Within-subjects)
One Independent Variable
Multiple Independent Variables
Correlation (Chapter 7)
Regression (Chapter 8)
One Sample z or t (Chapter 12)
One Sample t (Chapter 14)
Mann-Whitney U (Chapter 21)
Two Dependent Samples t (Chapter 15)
Sign Test or Wilcoxin Signed-Ranks Test (Chapter 21)
One-Way ANOVA (Chapter 18)
Kruskal-Wallis Test (Chapter 21)
Factorial ANOVA (Chapter 19)
Repeated Measures ANOVA (Chapter 18)
Friedman's Rank Test (Chapter 21)
Mixed Model ANOVA (Chapter 19)
Qualitative
Number of Variables
One Qualitative Variable
Two Qualitative Variables
Goodness-of-Fit (Chapter 20)
Test of Independence (Chapter 20)

FIGURE E.2 Flowchart summarizing various considerations when choosing a statistical procedure.

APPENDIX A

Review of Basic Mathematics

A.1 Introduction

This appendix offers information about basic skills that are useful in an introductory course in statistics. It is not intended to be a comprehensive review, nor should it be considered an initial unit of instruction for those who have no knowledge of the subject. It is intended primarily as a reminder of principles formerly learned but possibly covered with mental cobwebs.

A.2 Symbols and Their Meanings

Symbol	**Meaning**
$X \neq Y$	X is not equal to Y.
$X \sim Y$ or $X \approx Y$	X is approximately equal to Y.
$X > Y$	X is greater than Y.
$X < Y$	X is less than Y.
$X \geq Y$	X is equal to or greater than Y.
$X \pm Y$	As used in this book, it always identifies two limits: $X + Y$ and $X - Y$.

THE FAR SIDE®
By GARY LARSON
© 1982 FarWorks, Inc. All Rights Reserved/Dist. by Creators Syndicate
Larson

Early stages of math anxiety

The Far Side® by Gary Larson © 1982 FarWorks, Inc. All Rights Reserved. The Far Side® and the Larson® signature are registered trademarks of FarWorks, Inc. Used with permission.

Symbol	**Meaning**
$\lvert X \rvert$	The *magnitude* of X without regard to its sign; e.g., $\lvert +3 \rvert = 3$ and $\lvert -3 \rvert = 3$.
XY or $(X)(Y)$	The product of X and Y; X times Y.
$\frac{X}{Y}$ or X/Y	Alternative ways of indicating X divided by Y.
$\frac{Y}{X}$	The *reciprocal* of $\frac{X}{Y}$.
$\frac{1}{Y}$	The *reciprocal* of Y $\left(\text{reciprocal of } \frac{Y}{1}\right)$.
$(X)\left(\frac{1}{Y}\right)$	The product of X and the reciprocal of Y; an alternative way of writing X/Y.
$(XY)^2$	The square of the product of X and Y.
X^2Y^2	The product of X^2 and Y^2; it is the same as $(XY)^2$.
XY^2	The product of X and Y^2; the *"square" sign modifies Y but not X.*
∞	Infinity; a number indefinitely large.
4 or +4	When a *specific* number is written without a sign in front of it, a positive number is intended. Negative numbers are so indicated, e.g., −4.

A.3 Arithmetic Operations Involving Positive and Negative Numbers

Problem	**Comment**
$3 - 12 = -9$	To subtract a larger number from a smaller one, subtract the smaller from the larger and reverse the sign.
$3 + (-12) = -9$	Adding a negative number is the same as subtracting the number.
$3 - (-12) = 15$	Subtracting a negative number is the same as adding it.
$-3 - 12 = -15$	The sum of two negative numbers is the negative sum of the two numbers.
$(3)(-12) = -36$	The product of two numbers is negative when *one* of the two is negative.
$(-3)(-12) = 36$	The product of two numbers is positive when *both* are negative.
$(-2)^2 = 4$	The square of a negative number is positive, since to square is to multiply a number by itself.
$(-2)(3)(-4) = 24$	The product of more than two numbers is obtained by finding the product of any two of them, multiplying that product by one of the remaining numbers, and continuing this process as needed. Thus, $(-2)(3) = -6$ and $(-6)(-4) = 24$.
$(2)(0)(4) = 0$	The product of several terms is zero if any one of them is zero.
$2 + 3(-4) = 2 - 12 = -10$	In an additive sequence, reduce each term before summing. In the example, obtain the product *first*, then add it to the other term.
$\frac{-4}{2} = -2$	When *one* of the numbers in a fraction is negative, the quotient is negative.
$\frac{-4}{-2} = 2$	When *both* numbers in a fraction are negative, the quotient is positive.

A.4 Squares and Square Roots

Problem	**Comment**
$[(2)(3)(4)]^2 = (2^2)(3^2)(4^2)$ $24^2 = (4)(9)(16)$ $576 = 576$	The square of a product equals the product of the squares.
$(2 + 3 + 4)^2 \neq 2^2 + 3^2 + 4^2$ $9^2 \neq 4 + 9 + 16$ $81 \neq 29$	The square of a sum does *not* equal the sum of the squares.

Problem	**Comment**
$\left(\frac{4}{16}\right)^2 = \frac{4^2}{16^2}$ $\left(\frac{1}{4}\right)^2 = \frac{16}{256}$ $1/16 = 1/16$	The square of a fraction equals the fraction of the squares.
$\sqrt{(4)(9)(16)} = \sqrt{4}\,\sqrt{9}\,\sqrt{16}$ $\sqrt{576} = (2)(3)(4)$ $24 = 24$	The square root of a product equals the product of the square roots.
$\sqrt{9 + 16} \neq \sqrt{9} + \sqrt{16}$ $\sqrt{25} \neq 3 + 4$ $5 \neq 7$	The square root of a sum does not equal the sum of the square roots.
$\sqrt{\frac{4}{16}} = \frac{\sqrt{4}}{\sqrt{16}}$ $\sqrt{\frac{1}{4}} = \frac{2}{4}$ $\frac{1}{2} = \frac{1}{2}$	The square root of a fraction equals the fraction of the square roots.
$(\sqrt{4})^2 = 4$ $2^2 = 4$ $4 = 4$	The square of a square root is the same quantity found under the square root sign. Another example: $\sqrt{(x^2 - c)^2} = x^2 - c$.

A.5 Fractions

Problem	**Comment**
$\frac{1}{4} = .25$	To convert the ratio of two numbers to a decimal fraction, divide the numerator by the denominator.
$.25 = (100)(.25)\%$ $= 25\%$	To convert a decimal fraction to a percent, multiply by 100.
$\left(\frac{3}{5}\right)(16) = \frac{(3)(16)}{5}$ $= \frac{48}{5}$ $= 9.6$	To multiply a quantity by a fraction, multiply the quantity by the numerator of the fraction and divide that product by the denominator of the fraction.
$\frac{16}{4} = \left(\frac{1}{4}\right)(16) = 4$	To divide by a number is the same as multiplying by its reciprocal.

Problem	**Comment**
$\dfrac{16}{\frac{4}{5}} = \left(\dfrac{5}{4}\right)(16) = 20$	To divide by a fraction, multiply by its reciprocal.
$\dfrac{3+4-2}{8} = \dfrac{3}{8} + \dfrac{3}{4} - \dfrac{2}{8}$ $= \dfrac{5}{8}$	When the numerator of a fraction is a sum, the numerator may be separated into component additive parts, each divided by the denominator.
$\dfrac{3}{8} + \dfrac{4}{8} - \dfrac{2}{8} = \dfrac{3+4-2}{8}$ $= \dfrac{5}{8}$	When the several terms of a sum are fractions having a common denominator, the sum may be expressed as the sum of the numerators divided by the common denominator.
$\dfrac{(3)(15)}{5} = \dfrac{(3)(3)(\cancel{5})}{\cancel{5}}$ $= 9$	When the numerator and/or denominator of a fraction is the product of two or more terms, identical terms appearing in the numerator and denominator may be canceled.
$\left(\dfrac{1}{5}\right)\left(\dfrac{2}{7}\right)\left(\dfrac{3}{11}\right) = \dfrac{(1)(2)(3)}{(5)(7)(11)}$ $= \dfrac{6}{385}$	The product of several fractions equals the product of the numerators divided by the product of the denominators.

A.6 Operations Involving Parentheses

Problem	**Comment**
$2 + (4 - 3 + 2)$ $= 2 + 4 - 3 + 2$ $= 5$	When a positive sign precedes parentheses, the parentheses may be removed without changing the signs of the terms within.
$2 - (4 - 3 + 2)$ $= 2 - 4 + 3 - 2$ $= -1$	When a negative sign precedes parentheses, they may be removed if the signs of the terms within are reversed.
$a(b + c) = ab + ac$ A numerical example: $2(3 + 4) = 2(3) + 2(4)$ $2(7) = 6 + 8$ $14 = 14$	When a quantity within parentheses is to be multiplied by a number, *each* term within the parentheses must be so multiplied.
$2a + 4ab^2 = (2a)(1) + (2a)(2b^2)$ $= 2a(1 + 2b^2)$ A numerical example: $(6 + 8) = (2)(3) + (2)(4)$ $14 = 2(3 + 4)$ $14 = (2)(7) = 14$	When all terms of a sum contain a common multiplier, that multiplier may be factored out as a multiplier of the remaining sum.

Problem	Comment
$3 + (1 + 2)^2 = 3 + 3^2$ $= 3 + 9 = 12$	When parentheses are modified by squaring or some other function, take account of the modifier before combining with other terms.
$\left[100 - 40\left(\frac{20}{10}\right)\right] + \left[\frac{20}{10} + (40 - 30)\right]$ $= [100 - 40(2)] \quad + [2 + 10]$ $= [100 - 80] \quad + 12$ $= 20 \quad + 12$ $= 32$	When an expression contains nesting parentheses, *perform those operations required to remove the most interior parentheses first*. Simplify the expression by working outward.
$(a + b)(c - d) = ac - ad + bc - bd$ A numerical example: $(2 + 3)(5 - 4)$ $= (2)(5) + (2)(-4) + (3)(5) + (3)(-4)$ $= 10 - 8 + 15 - 12$ $= 5$	The product of two sums may be obtained by multiplying each element of one term by each element of the other term.
$(a - b)^2 = a^2 - 2ab + b^2$ A numerical example: $(2 - 4)^2 = 2^2 + 2(2)(-4) + (-4)^2$ $(-2)^2 = 4 - 16 + 16$ $4 = 4$	The square of a binomial equals the square of the first term plus two times the product of the two terms plus the square of the second term.

A.7 Equations in One Unknown

To solve an equation in an unknown, we must work to isolate the unknown, unmodified and by itself, on one side of the equation, and everything else on the other side of the equation. In working toward this goal, one side of the equation may be altered if the other side is altered in the *same* way. Alteration by adding, subtracting, multiplying, dividing, squaring, taking the square root, or taking the reciprocal may each be useful, depending on circumstances. *In each of the following problems, we shall suppose that it is required to solve for* Y.

Problem	Comment
$Y + 3 = 4$ *Operation:* subtract 3 from both sides $Y = 1$	Subtracting 3 isolates Y on one side of the equation.
$3Y = 12$ *Operation:* divide both sides by 3 $Y = 4$	Dividing by 3 isolates Y.
$\frac{Y}{3} = 4$ *Operation:* multiply both sides by 3 $Y = 12$	Multiplying by 3 isolates Y.

Many problems require two or more successive operations to isolate the unknown. Several examples follow. The order of operations is generally important, although sometimes there is more than one solution.

Problem	**Comment**
$\frac{3}{Y} = 4$ *First operation:* take the reciprocal of both sides $\frac{Y}{3} = \frac{1}{4}$ *Second operation:* multiply both sides by 3 $Y = \frac{3}{4}$	Y must be dug out of the denominator, and taking the reciprocal is one way to do it. Can you think of another way?
$2Y - 4 = 8$ *First operation:* add 4 to both sides $2Y = 12$ *Second operation:* divide both sides by 2 $Y = 6$	In this problem, the two operations could be performed in reverse order with no loss of efficiency.
$2(4 - Y) = 12$ *First operation:* perform the indicated multiplication $8 - 2Y = 12$ *Second operation:* subtract 8 from both sides $-2Y = 4$ *Third operation:* divide both sides by -2 $Y = -2$	There is another way to isolate Y; can you think of it?
$5Y - 3 = Y + 5$ *First operation:* subtract Y from both sides $4Y - 3 = 5$ *Second operation:* add 3 to both sides $4Y = 8$ *Third operation:* divide both sides by 4 $Y = 2$	Like terms must be collected in solving an equation.

A.8 Summation Rules

In the algebraic transformation of statistical equations, there are three basic rules of summation that come into frequent play. The rules are as follows:

Rule 1: *When a quantity that is itself a sum or difference is to be summed, the summation sign may be distributed among the separate terms of the sum*. That is,

$$\sum(X + Y) = \sum X + \sum Y$$

Rule 2: *The sum of a constant equals the product of the constant and the number of times the constant appears.* That is,

$$\sum a = na$$

Rule 3: *The sum of the product of a constant and a variable is equivalent to the product of the constant and the sum of the variable.* That is,

$$\sum aX = a\sum X$$

A.9 Test of Mathematical Skills

Now that you have reviewed, see how many of the following questions you can answer correctly.

Reference for Questions 1–4: Section A.2.

1. What symbol, inserted between 3 and 2, indicates that 3 is greater than 2? ________
2. The reciprocal of 2/3 is ________
3. $|-8| = ?$ ________
4. Write $(ab)^2$ another way ________

Reference for Questions 5–11: Section A.3.

5. $2 - (-4) = ?$ ________
6. $(3)(-4) = ?$ ________
7. $(-3)(-4) = ?$ ________
8. $(2)(-3)(-2)(-2) = ?$ ________
9. $(2)(-3)(0)(-1) = ?$ ________
10. $2 - 4(3/4) = ?$ ________
11. $-4/-2 = ?$ ________

Reference for Questions 12–18: Section A.4.

12. Does $(abc)^2 = abc^2$? ________
13. Does $a^2 + b^2 = (a + b)^2$? ________
14. Does $\dfrac{a^2}{b^2} = \left(\dfrac{a}{b}\right)^2$? ________
15. Does $\sqrt{4b} = 2\sqrt{b}$? ________
16. Does $\sqrt{a^2 + b^2} = a + b$? ________
17. Does $\sqrt{\dfrac{a^2}{b^2}} = \dfrac{a}{b}$? ________
18. The square of $\sqrt{a} = ?$ ________

Reference for Questions 19–23: Section A.5.

19. $3/12 =$ what percent? ________
20. Simplify: $\dfrac{2ab^2}{a}$ ________
21. Simplify: $\dfrac{a}{\dfrac{ab}{c}}$ ________

22. Does $\dfrac{a-b}{c} = \dfrac{a}{c} - \dfrac{b}{c}$? ______

23. Does $12/15 = 4/5$? ______

Reference for Questions 24–29: Section A.6.

24. $a - (b - a) = ?$ ______

25. $a(X + Y) = ?$ ______

26. $2ab^2 - 2b$ may be expressed as the product of what two quantities? ______

27. $(X + 4)(Y - 2) = ?$ ______

28. $(X - Y)^2 = ?$ ______

29. $\left\{20 - 3\left[3 + \left(\dfrac{12}{4}\right)\right]\right\}^2 = ?$ ______

Reference for Questions 30–35: Section A.7.

30. $Y - 3 = 2$ $Y =$ ______

31. $2Y = -6$ $Y =$ ______

32. $(3/4)Y = 12$ $Y =$ ______

33. $\dfrac{2}{Y} = 3$ $Y =$ ______

34. $2(Y - 2) = 4$ $Y =$ ______

35. $3Y - 3 = 2Y + 5$ $Y =$ ______

A.10 Answers to Test Questions

1. $>$
2. 3/2
3. 8
4. a^2b^2
5. 6
6. -12
7. 12
8. -24
9. Zero
10. -1
11. 2
12. No
13. No
14. Yes
15. Yes
16. No
17. Yes
18. a
19. 25
20. $2b^2$
21. c/b
22. Yes
23. Yes
24. $2a - b$
25. $aX + aY$
26. $2b$ and $(ab - 1)$
27. $XY + 4Y - 2X - 8$
28. $X^2 - 2XY + Y^2$
29. 4
30. 5
31. -3
32. 16
33. 2/3
34. 4
35. 8

APPENDIX B

List of Symbols

Symbols used in this book are identified below. Numbers in parentheses indicate the page on which each symbol is introduced and defined and where pronunciation is given for Greek letters.

Greek Letter Symbols

α	level of significance, probability of a Type I error (186)
β	probability of a Type II error (220)
$1 - \beta$	power of a test (220)
E^2_R	epsilon-squared (397)
η^2	eta-squared, a measure of effect size for F (316)
μ, μ_X	mean of a population (51)
$\mu_{\overline{X}}$	mean of the sampling distribution of means (175)
$\mu_{\overline{X}-\overline{Y}}$	mean of the sampling distribution of $\overline{X} - \overline{Y}$ (234)

Ω	odds (163)
θ	population odds ratio (376)
ρ_{XY}	population Pearson correlation coefficient (275)
Σ	the sum of (51)
σ, σ_X	standard deviation of a population (66)
σ^2, σ_X^2	variance of a population (65)
σ_r	standard error of r (275)
$\sigma_{\overline{X}}$	standard error of the mean (175)
$\sigma_{\overline{X}-\overline{Y}}$	standard error of the difference between two means (236)
ϕ	phi coefficient (376), Cramer's phi (375)
χ^2	chi-square (366)
χ^2_F	test statistic for Friedman's rank test (397)
$\hat{\omega}$	measure of effect size for χ^2 (369)
$\hat{\omega}^2$	omega-squared, a measure of effect size for F (316)

English Letter Symbols

a	Y intercept of a line (144)
b	slope of a line (144)
C	confidence coefficient (285)
D	differences between paired scores (261)
$\overline{D}$	mean difference between paired scores (261)
d	(1) discrepancy between a hypothesized value and the true value; (2) effect size (217)
df	degrees of freedom (197)
F	variance ratio (306)
f	frequency (21)
f_e	expected frequency (365)
f_o	observed frequency (365)
g	effect size (217)
H	test statistic for the Kruskal–Wallis test (395)
H_A	alternative hypothesis (185)
H_0	null hypothesis (185)
i	class interval width (21)
K	a comparison (322)
k	number of groups (352)
k^2	coefficient of nondetermination (148)
Mdn	median (49)
Mo	mode (49)

MS	mean square, or variance estimate (306)
N	number of cases in a population (21)
n	number of cases in a sample (21)
n_i	number of cases in the *i*th subgroup (303)
$\tilde{n}$	harmonic mean (319)
OR	odds ratio of a sample (376)
P	population value of a proportion (364)
P_j	*j*th percentile (26)
Pr	probability (156)
q	Studentized range statistic (319)
Q	semi-interquartile range (63)
R_+, R_-	sum of the ranks with a positive or a negative sign (394)
r	(1) sample Pearson's correlation coefficient (103); (2) effect size (249)
r^2	coefficient of determination (151)
r_C	matched-pairs rank biserial correlation coefficient (151)
r_g	Glass rank biserial correlation coefficient (389)
r_{S}	Spearman's rank-order correlation coefficient (112)
S, S_X	standard deviation of a sample (65)
S^2, S_X^2	variance of a sample (65)
$S_{Y'}^2$	variance of predicted scores (165)
S_{YX}	standard error of estimate (132)
S_{YX}^2	variance of obtained scores about their predicted scores (151)
s, s_X	estimate of a population standard deviation (194)
s^2, s_X^2	unbiased estimate of a population variance (194)
s_{bet}^2	between-groups variance estimate (306)
s_C^2	column variance estimate (347)
$s_{\overline{D}}$	estimate of the standard error of the mean of difference scores (262)
s_k	standard error of a comparison (320)
s_p^2	pooled estimate of the population variance (238)
s_{resid}^2	residual variance estimate (325)
s_R^2	row variance estimate (346)
$s_{R\times C}^2$	row by column interaction variance estimate (347)
s_{w}^2	within-groups variance estimate (305)
s_{wc}^2	within-cells variance estimate (346)
$s_{\overline{X}}$	estimate of the standard error of the mean (194)
$s_{\overline{X}-\overline{Y}}$	estimate of the standard error of the difference between two means (237)
SS	sum of squares (65)
T	test statistic for the Wilcoxon signed-ranks test (394)
t	Student's *t* statistic (195)

t_p	t beyond which a given proportion of values fall (284)
U	test statistic for the Mann-Whitney test (387)
W	concordance coefficient (399)
w	maximum desired difference between sample and population characteristics in an interval estimate (294)
X	raw scores (51)
$\overline{X}$	mean of a sample (51)
$\overline{\overline{X}}$	mean of all scores, or grand mean (303)
Y	raw scores (51)
$\overline{Y}$	mean of a sample (51)
Y'	predicted raw score (127)
z	z score, or standard score (75)
z'_Y	predicted z score (129)
z'	Fisher's logarithmic transformation of r (277)

APPENDIX C

Answers to Problems

CHAPTER 1

1. (a) First (b) Both (c) Second
(d) First (e) First

3. (a) Inferential (b) Inferential
(c) Descriptive (d) Descriptive

5. Yes; a statistical conclusion is a conclusion about the data only. A research conclusion derives partly from the statistical conclusion, but other factors must also be considered, such as the adequacy of the experimental design. As an example, review the executive monkey experiment (Section 1.3).

7. (a) Continuous (b) Continuous
(c) Discrete (d) Discrete (e) Discrete
(f) Continuous (the underlying variable is continuous) (g) Continuous
(h) Continuous

9. (a) No; the accuracy of measurement was only to the nearest tenth of a pound. (b) Yes (c) No; the investigators must have been interested in weight to the nearest tenth of a pound; otherwise, they would not have measured with that degree of accuracy.

11. Ordinal

13. (a) Yes (b) Yes; in each case, a given interval represents the same difference no matter where it is located along the scale.

15. (a) Something between ordinal and interval; greater than ordinal because a difference of 2 points probably reflects a greater difference in knowledge than a difference of 1 point, yet less than interval because a difference of 5 points at one part of the scale may not reflect the same as a 5-point difference at another part of the scale. (b) No; Bob may be able to answer a question about American history that was not included on the test. (c) Not necessarily; the test is not on a ratio scale (there is no meaningful zero point). (d) No; the test may have completely tapped Mark's knowledge of American history, but Sheela may have additional knowledge, or vice versa.

CHAPTER 2

1. (a)

FAVORITE FOREIGN FOOD	FREQUENCY
Chinese	9
Mexican	6
Italian	4
Greek	2
Indian	2
Japanese	1

(b) Nominal (c) No

3. (a) 51.5–52.5 (b) 750–850 yds.
(c) 455–465 lb. (d) .55–.65 in.
(e) .465–.475 sec.

5. (a) Real limits are determined by taking 1/2 the smallest unit of measurement (in this case, whole numbers), so a score of 39.5 is not possible. (b) 30–39 and 40–49, $i = 10$

7. (Answers for interval width and apparent limits may differ from those given, depending on the width of class interval selected.)
(a) range = 39, i = 3, 36–38, 75–77
(b) range = 63, i = 5, 50–54, 115–119
(c) range = 144, i = 10, 20–29, 170–179
(d) range = 43, i = 3, −24 to −22, +21 to +23
(e) range = 2.34, i = .20, 1.00–1.19, 3.40– 3.59
(f) range = 534, i = 50, 250–299, 800–849

9. No; if you make the lower apparent limits of each interval a multiple of the interval width (guideline #7), you will arrive at the same score intervals.

11. (a) .37 (b) .563 (c) .042
(d) .0921 (e) .004 (f) .0085
(g) .0002

13. (a)

	Rel *f* (%)	
SCORE	*Good mood (A)*	*Bad mood (B)*
155–159		2.0
150–154	1.33	4.0
145–149	2.67	14.0
140–144	4.67	24.0
135–139	8.00	20.0
130–134	9.33	14.0
125–129	16.67	8.0
120–124	15.33	6.0
115–119	12.00	0
110–114	13.33	4.0
105–109	8.00	2.0
100–104	5.33	0
95–99	2.00	2.0
90–94	1.33	
	99.99	100.0

(b) Relative frequency distributions are more meaningful for comparing two distributions in which the number of cases differ because they put both distributions on the same basis.

15. (a) 74.5 (b) 83.5 (c) 68.5
(d) 89.5 (e) 4.0 % (f) 32.0 %
(g) 98.0 % (h) 0

17. (a) 91.33

(b) $$100\left[\frac{86 + \left(\frac{1.5}{5} \times 25\right)}{150}\right] = 62.3$$

(c) $.25n = 37.5$

$$\frac{12.5}{20} \times 5 = 3.125$$

$$P_{25} = 109.5 + 3.125 = 112.625$$

(d) 122.11

19. (a) Range = 99 − 52 = 47

$$\left.\begin{array}{l}47/10 = 4.7\\ 47/20 = 2.35\end{array}\right\}\quad i = 3 \text{ or } 4$$

SCORE	(a) FREQUENCY	(b) CUM f	(c) CUM %
97–100	2	50	100
93–96	2	48	96
89–92	5	46	92
85–88	7	41	82
81–84	9	34	68
77–80	5	25	50
73–76	4	20	40
69–72	8	16	32
65–68	2	8	16
61–64	2	6	12
57–60	2	4	8
53–56	1	2	4
49–52	1	1	2
	n = 50		

Note: For this example, we chose to ignore guideline #7 (Section 2.3) and have 100 as our upper apparent limit for the highest interval. We could have begun with 96–99, but that would have excluded the highest possible score of 100. A frequency distribution with i = 3 was also possible.

CHAPTER 3

1.

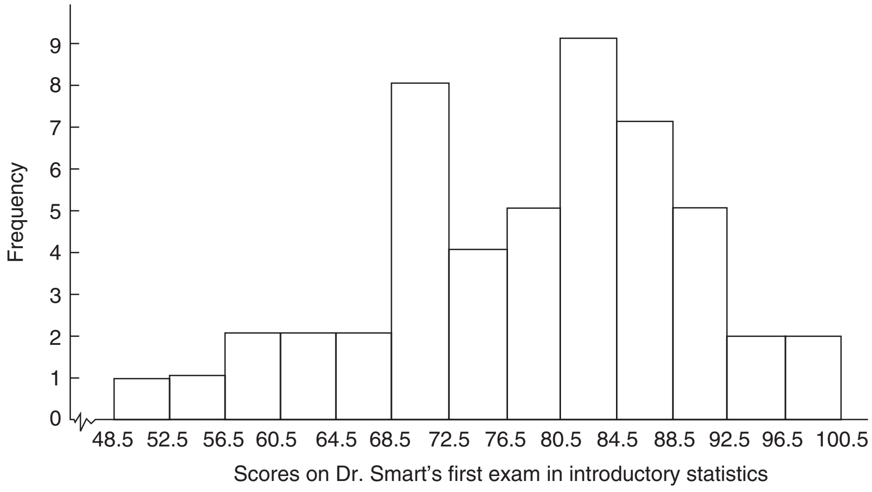
	LEFT	RIGHT
(a)	56.5	59.5
(b)	65.5	68.5
(c)	74.5	77.5
(d)	83.5	86.5
(e)	92.5	95.5

3. (a) 61 (b) 70 (c) 79 (d) 88 (e) 97

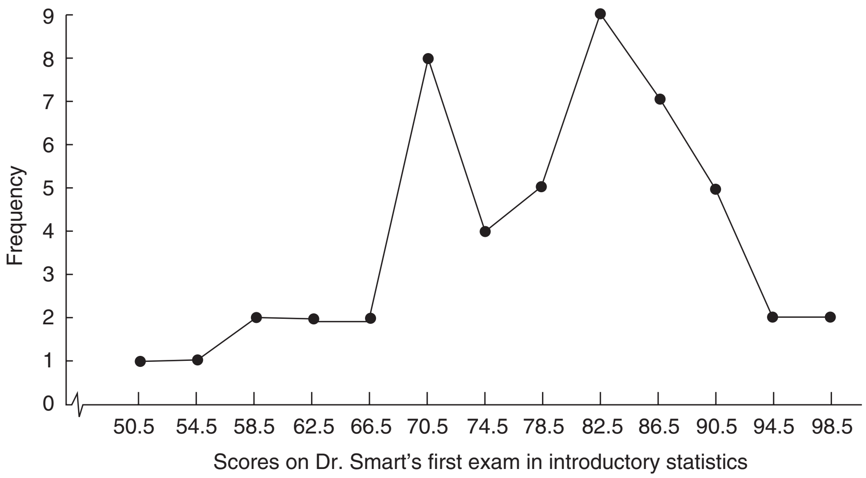
5. Ordinarily, we bring the polygon down to the horizontal axis at both ends. However, there are problems when scores in the next adjacent class interval are not possible, as in this case. Bringing the polygon down to the midpoint of the next adjacent interval may mislead someone looking at the graph to think that scores in that interval were possible. Bringing the polygon straight down implies that there were scores at two frequencies, including f = 0. The best thing to do in a case like this is to leave the dot "dangling."

7. Using the same intervals that we used for Problem 19 in Chapter 2 (the highest interval was 97–100), we get:

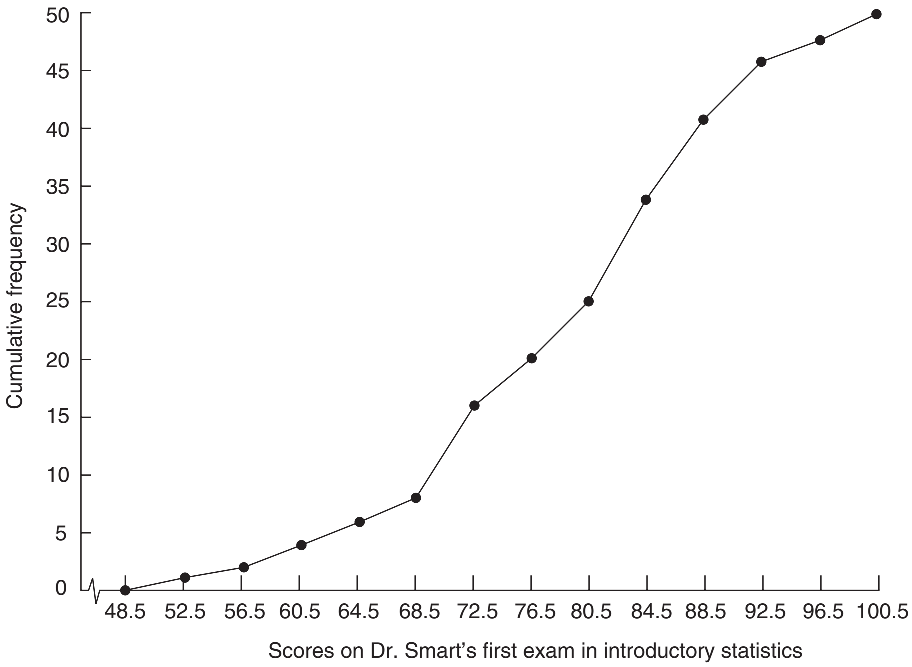
(a)



(b)

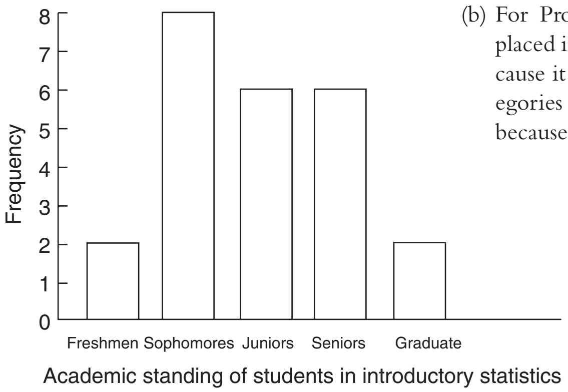


(c)



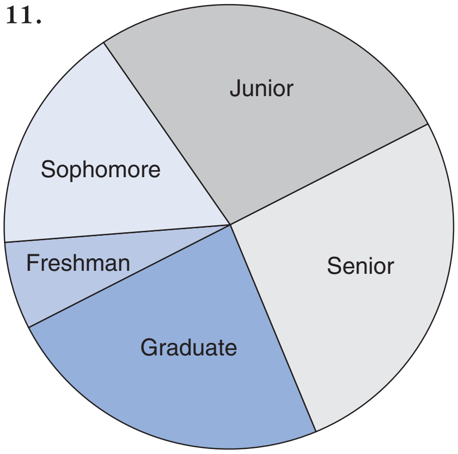
(d) Because the midpoints of the intervals would have been whole numbers.

9. (a)



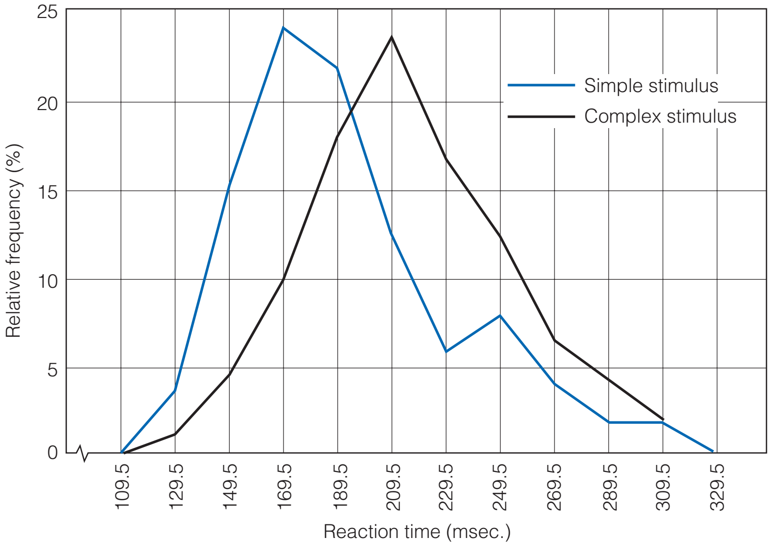
(b) For Problem 8, the categories may be placed in any order along the abscissa because it is a nominal scale. Here, the categories must be placed in order of rank because it is an ordinal scale.

11.



Male psychology majors

13.

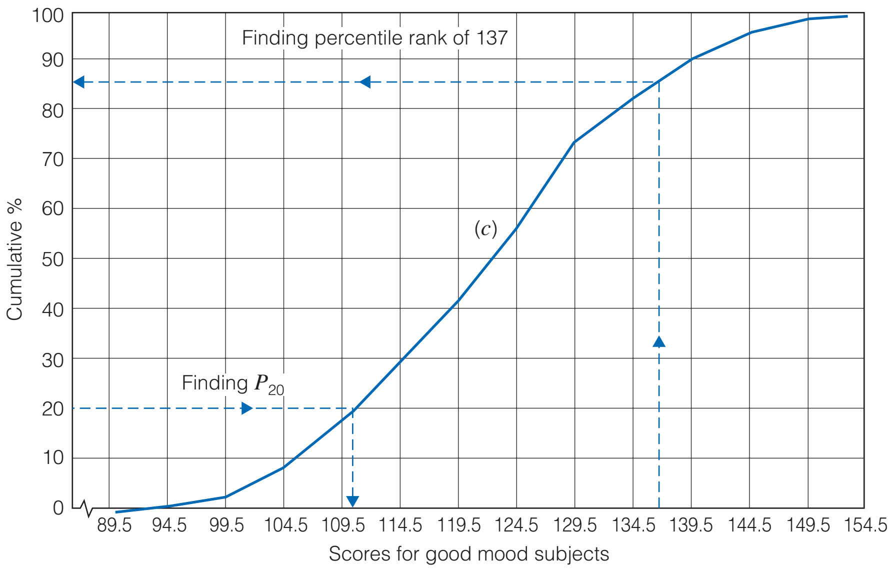


Conclusion: The spread of reaction times under the two conditions is similar, but a greater proportion of subjects react more quickly to the simple stimulus.

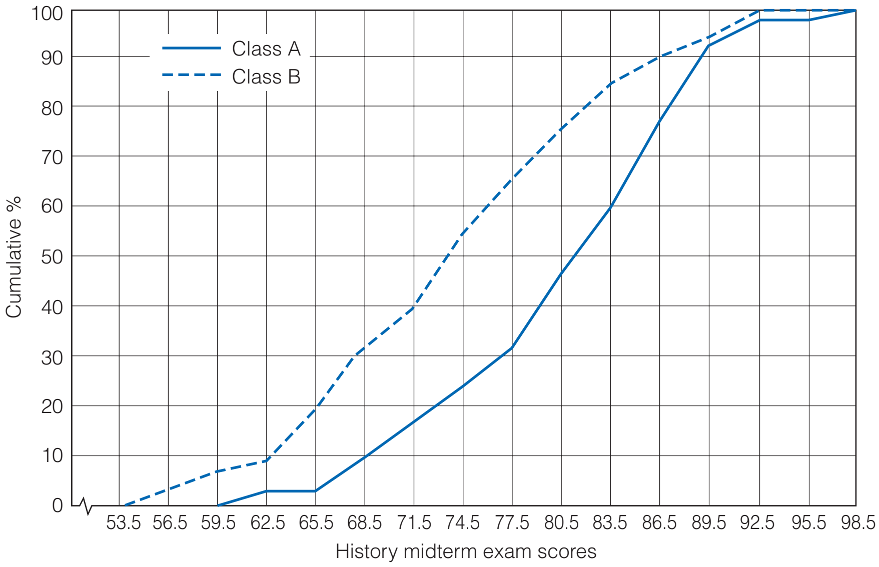
15.

SCORE FOR GOOD MOOD SUBJECTS	FREQUENCY	(a) CUM *f*	(b) CUM %
150–154	2	150	100.0
145–149	4	148	98.7
140–144	7	144	96.0
135–139	12	137	91.3
130–134	14	125	83.3
125–129	25	111	74.0
120–124	23	86	57.3
115–119	18	63	42.0
110–114	20	45	30.0
105–109	12	25	16.7
100–104	8	13	8.7
95–99	3	5	3.3
90–94	2	2	1.3

(c–d)



17. (a)



(b) Distribution A falls to the right of Distribution B. (c) The horizontal space occupied by Distribution B is greater. (d) The cumulative curves are smoother. (e) The noncumulated distribution shows the frequency of a particular score or range of scores; the cumulative distribution shows the frequency of scores below a particular score point and is better for identifying the position of a score relative to the group.

CHAPTER 4

1. (a) Median = 12.5; mean = 12.25 (b) Median = 12; mean = 12 (c) Median = 11; mean = 11 (d) Median = 10; mean = 9.67 (e) Median = 8; mean = 7.17 (f) Median = 6; mean = 6.5 (g) Median = 7.17; mean = 6.5 (h) Median = 3.83; mean = 3.5 (i) Median = 2; mean = 2.62

3. (a) Mean (b) Mode (c) Mean (d) Median (e) Mean (f) Median (g) Mean (h) Mean

5. Yes; the median is 4 because it falls between the fifth and sixth scores (3 and 5).

7.

$$\overline{X} = \frac{\Sigma X}{n}$$

$$4.8 = \frac{41+X}{10}$$

$$48 = 41 + X$$

$$7 = X$$

9. In this case, "average" refers to the mean.

11. The SAT scores of the 40 National Merit scholars probably increased the value of the mean SAT score (remember, the mean is responsive to the exact position of each score), but had no effect on the median.

13. The median; it is a reasonably stable measure, but not as responsive as the mean to the few deviant scores.

15. (a) $Mo = 12$ (b) $Mdn = 13$ (c) $\overline{X} = 13.3$ (d) Only the mean changed because it is responsive to each score in the distribution.

17. (a) 25 (b) 5 (c) 100 (d) 5 (e) 20 (f) 16

19. Yes; in popular usage, the term *average* is often used indiscriminately for any measure of central tendency. If one organization uses one index and the other organization uses another index, their statements could be "consistent," assuming the distribution of wages is skewed.

CHAPTER 5

1. (a) Semi-interquartile range (b) Range (c) Standard deviation (d) Range (e) Semi-interquartile range (f) Standard deviation (g) Semi-interquartile range

3. Simple stimulus: (a) 199, (b) 25.4 Complex stimulus: (a) 199, (b) 25.4 (c) Variability of the two in terms of range and Q is equal.

5. (a) 85 (b) 68 (c) 80

7. 15, 12, 11, 2, and 10

9. (a) $S^2 = 10.44$ (b) $S = 3.23$

11. (a) Inches^2 (b) Inches

13. (a) 9 (b) 3 (c) 9 (d) 3 (e–f) Because the Y scores are obtained by adding a constant (2) to each X score; adding a constant to each score in a distribution does not affect the variance or standard deviation.

15. (a) No; the variance is obtained by squaring scores and thus cannot have a negative value. (b) No; the standard deviation is the square root of the variance, which always is a positive value.

17. With open-ended distributions, the semi-interquartile range may be the only measure that is reasonable to compute.

19. (a) 40,000 (b) 12,437

21. (a) +1.00 (b) +2.00 (c) −2.00 (d) 0.00 (e) −1.33 (f) −0.33 (g) +0.53 (h) −0.87 (i) +1.47

23. Nothing; a score is uninterpretable without a frame of reference (such as the mean and standard deviation).

25. Robin; her score is 1.5 z scores (standard deviations) greater than the mean for her exam, whereas Susan's score is 1.17 z scores greater than the mean for her exam.

27. −1.11, −1.51, +1.71, +0.91, +0.50, +0.10, +0.10, and −0.71. The shape of the distribution has not changed because to calculate each z score, we first subtract a constant (the mean) and then divide by another constant (the standard deviation).

29. In IBM SPSS, the mean is 79.92 and the standard deviation is 7.93298. When using n in the denominator, the standard deviation is

$$\sqrt{(7.93298)^2\left(\frac{49}{50}\right)} = 7.85325.$$

CHAPTER 6

1. The normal curve is a mathematical abstraction having a particular defining equation. No real variable is exactly normally distributed, but the normal curve does offer a convenient and reasonably accurate *description* of a number of distributions (especially the hypothetical distributions that we will encounter in inferential statistics).

3. (a) .8413 (b) .50 (c) .9772 (d) .8023

5. (a) .2957 (b) .1359 (c) .0215

7. (a) .3085 (b) .1587 (c) .2119 (d) .9332 (e) .3023 (f) .8716 (g) .0934 (h) .3056

9. (a) +1.64 (b) +1.96 (c) +2.33 (d) +2.58

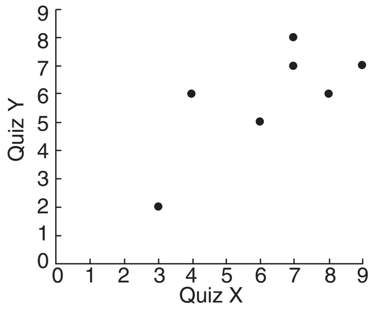
11. .8944

13. 31.68 and 48.32 (z scores of ±1.04)

15. (a) About .0004 (b) 72,000

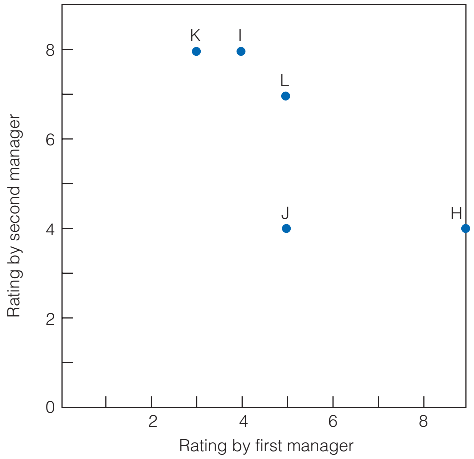
CHAPTER 7

1.



(a) Positive (b) $\Sigma X = 44$, $\Sigma Y = 41$, $\Sigma X^2 = 304$, $\Sigma Y^2 = 263$, $\Sigma XY = 276$, $r = .73$ (c) Consistency

3. (a)



The relationship is negative.
(b) $\Sigma X = 26$; $\Sigma Y = 31$; $\Sigma X^2 = 156$; $\Sigma Y^2 = 209$; $\Sigma XY = 147$; $r = -.76$
(c) Agreement

5. To see the parallels between the two formulas, start with the formula for $\Sigma(X-\overline{X})^2$ and write out in full all expressions involving the superscript 2. Then substitute Y for the second X in any such expressions.

Specifically:

$$\begin{aligned} \Sigma(X-\overline{X})^2 &= \Sigma(X-\overline{X})\,(X-\overline{X}) \\ &= \Sigma X^2-(\Sigma X)^2/n \\ &= \Sigma XX-(\Sigma X)\,(\Sigma X)/n \\ \text{then}\quad & \Sigma(X-\overline{X})\,(Y-\overline{Y}) \\ &= \Sigma XY-(\Sigma X)\,(\Sigma Y)/n \end{aligned}$$

7. (a) 1, 2, 3, 4.5, 4.5; $\Sigma R = 5(5+1)/2 = 15$
(b) 1, 3, 3, 3, 5, 6.5, 6.5, 8;
$\Sigma R = 8(8+1)/2 = 36$

9. Yes; the scores are in the form of ranks and there are no ties.

11. Start with a set of original (untransformed) scores. For each, find the transformed score according to any of the so-called linear transformations. In a two-dimensional plot, where the X axis shows the original and the Y axis the transformed score, each point whose coordinates are an original and the corresponding transformed score will lie on the same straight line.

13. (a) Positive (b) A correlation between two variables cannot by itself prove a causal relationship between them. (c) A third variable or a complex of additional variables may influence both smoking and lung cancer, causing them to covary. For example, certain temperaments may predispose people both to smoke and to develop lung cancer. Or the third variable may be socioeconomic status; lower-class people are generally heavier smokers than middle-class people, and they generally live where air pollution is worse. (d) Besides the simple correlational findings, other evidence includes sophisticated correlational findings in which third variables are ruled out as alternative explanations; experiments in which laboratory animals are exposed to tobacco smoke; and biological research showing exactly how smoking damages the tissues and cells of the lungs.

15. (a) Positive (b) No; a correlation between two variables cannot by itself prove a causal relationship between them. (c) A third variable influences both the quantity of ice cream sold and the number of murders, causing them to covary: the temperature. On hotter days, people buy more ice cream, and on hotter days murders become more likely—perhaps because the heat brings extra irritability or because it raises testosterone levels. (d) "c"

17. There is no such thing as *the* correlation between two variables. At the Minuscule Corporation, the correlation may prove to be much lower.

19. Because all the persons in the group are highly creative, the range of talent on this variable is small. The two variables would probably correlate more substantially in a random sample of the general population.

21. For the sample of ninth and twelfth graders because of the discontinuities in the heights and in the weights resulting from the removal of the students in the intermediate grades.

CHAPTER 8

1. When the relationship between X and Y is linear, and when the data are a sample from a population rather than the population itself.

3. Zero. When $r = -1.00$, all data points lie on a (straight) line, and this is the line of best fit. Thus $Y' = Y$ for each case, $d_Y = 0$ for each case, $d_Y^2 = 0$ for each case, and $\Sigma d_Y^2 = 0$.

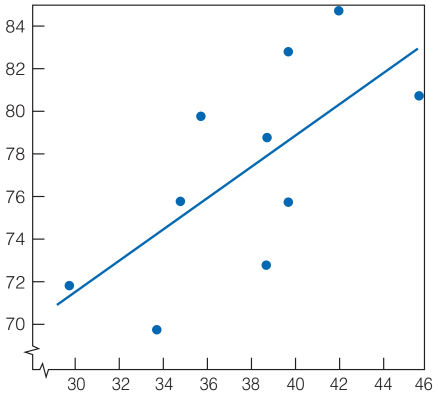
5.
(a) $z_X = 0.00$, $z'_Y = 0.00$, $Y' = 500$
(b) $z_X = 1.00$, $z'_Y = 0.50$, $Y' = 550$
(c) $z_X = 2.00$, $z'_Y = 1.00$, $Y' = 600$
(d) $z_X = -1.00$, $z'_Y = -0.50$, $Y' = 450$
(e) $z_X = -2.00$, $z'_Y = -1.00$, $Y' = 400$
(f) $z_X = 3.00$, $z'_Y = 1.50$, $Y' = 650$
(g) $z_X = -3.00$, $z'_Y = -1.50$, $Y' = 350$

7. No; here $X = \overline{X}$ and $Y' = \overline{Y}$, but no matter what the value of r, $Y' = \overline{Y}$ if $X = \overline{X}$.

9. The least-squares criterion calls for minimizing Σd_Y^2, which is $\Sigma(Y - Y')^2$. When $r = 0.00$, $Y' = \overline{Y}$ for all values of X, so $\Sigma d_Y^2 = \Sigma(Y - \overline{Y})^2$, and this sum is lower than the sum of the squared deviations about any quantity other than the mean of the Y scores.

11. $Y' = 3.333X + 166.667$ (a) For $X = 100$, $Y' = 3.333(100) + 166.667 = 500$ (c) For $X = 130$, $Y' = 3.333(130) + 166.667 = 433.290 + 166.667 = 600$ (e) For $X = 70$, $Y' = 3.333(70) + 166.667 = 233.310 + 166.667 = 400$

13. (a)



(b) $Y' = .746X + 49.06$ (c) See graph. (d) 70.0 (e) 77.4 (f) 84.1

15. (a) −77.74 lb. (b) The weight is impossible, and thus the regression equation fails to work for low values of X. It is valid only for adult heights like those in the data that generated the equation.

17. (a) 86.6 (b) 13.73

19. (a) $Y' = .00233X + 1.343$ (b) 2.475 and 3.000 (c) Linearity of regression (d) .30311 (e) $z = .58$ for the mean of 2.65; proportion = .2810 (f) $z = -1.57$ for a GPA of 2.0; proportion = .0582 (g) 1.88 and 3.07 (h) $z = -1.65$ for a GPA of 2.5; proportion = .9505 (i) $z = .00$ for a GPA of 3.0; proportion = .5000 (j) $z = \pm 0.67$; 2.80 and 3.20 (k) Linearity of regression, homoscedasticity, and normality of the distribution of Y values for both values of X.

21. (a) $Y' = .23X + 6.794$ (b) $S_{YX} = 4.707$

23. $Y' = .23X + 6.794$: (a) 9.09, (b) 11.39, (c) 13.69, (d) 15.99

CHAPTER 9

1. $S_{YX}^2 = S_Y^2(1 - r^2) = S_Y^2 - S_Y^2 r^2$

$$S_Y^2 r^2 = S_Y^2 - S_{YX}^2$$

$$r^2 = \frac{S_Y^2 - S_{YX}^2}{S_Y^2} = 1 - \frac{S_{YX}^2}{S_Y^2}$$

$$r = \sqrt{1 - \frac{S_{YX}^2}{S_Y^2}}$$

3. (a) No; this equation specifies a Y intercept (+5), but in standard-score form, a regression equation has no Y intercept. (b−c) It must be negative. Because standard deviations must be positive, the sign of the regression coefficient, (rS_Y/S_X), is determined by r. Here the regression coefficient is negative, (−.25), so r must be negative. (d) The correlation is negative, and on average, Y decreases by .25 unit for each unit of increase in X, but we cannot tell how strong the correlation is.

5. This equation predicts that for all values of X, Y is 12. Thus $r = 0$ and $\overline{Y} = 12$.

7. (a) 98 (b) The slope is 98, a very high value, so the line is very steep in sloping from lower left to upper right.

9. Still well above average, but lower than 150 because of regression on the mean.

11. (a) No, because of regression on the mean. (b) No, for the same reason.

13. (a) .54 (b) In predicting either variable from the other, the errors are 54% as great, on average, as those that would occur if there were no relationship between the variables. (c) .46 (d) Forty-six percent of the variation in one variable is associated with differences in the other variable.

15. $\Sigma(Y - \overline{Y})^2$, because when r is 0, for each case $Y' = \overline{Y}$.

17. When $Y' = \overline{Y}$ for each case, which occurs when $r = 0$.

19. (a) The predictive error relative to the worst possible case improves across these values of r from .99+ to .95 to .80 to .44. (b) The proportion of variation in common between the two variables increases across these values of r from .01 to .09 to .36 to .81. (c) The proportion of correct placements increases across these values of r from .53 to .60 to .70 to .85.

CHAPTER 10

1. (a) .25 (b) 25

3. (a) (1/6) × (1/6) = 1/36 (b) (1/6) × (1/6) = 1/36 (c) [(1/6) × (1/6)] + [(1/6) × (1/6)] = 2/36 = 1/18

5. (a) (1/52) × (1/52) = 1/2704
(b) (4/52) × (4/52) = 16/2704 = 1/169
(c) (13/52) × (13/52) = 169/2704 = 1/16 (d) (4/52) × (4/52) = 16/2704 = 1/169 (e) [(4/52) × (4/52)] + [(4/52) × (4/52)] = 32/2704 = 2/169

7. Empirical, because they are derived from experience in similar circumstances.

9. (a) .0156 (b) .0938 (c) .2344

11. (a) .2503 (b) .7759

13. (a) .0042

15.
$$\left(\frac{30}{3{,}000{,}000}\right) \times \left(\frac{30}{3{,}000{,}000}\right) = \frac{1}{10{,}000{,}000{,}000} = \frac{1}{10^{10}}$$

(one in 10 billion).

Diaconis and Mostellar calculated the probability that *any one* of the millions of people who enter two lotteries within 4 months will win both. The probability of this event is much higher because there are millions of chances for it to occur = one chance for each of the people who enter two lotteries. Your friend has only one of those chances.

CHAPTER 11

1. A person in the population will have an unfairly large chance to get into your sample if the person speaks fluent English, has an opinion on the issue, and is not in a hurry. Bias may also arise because you prefer people of a certain age or gender or ethnic group.

3. One should assign to each score a unique number; the conventional choices would be the two-digit integers 01–80. In the table of random numbers, one should arbitrarily select a starting location and a direction in which to read off successive numbers, two at a time. One should choose for the sample those scores designated by the first 10 selections from the table that lie within the range 01 to 80.

5. $\mu_X = 5.0$; $\sigma_X = 3.63$

(b) Samples	(c) $\overline{X}$	(d) $\overline{X}$	f
1,1	1.0	1.0	1
1,2	1.5	1.5	2
1,4	2.5	2.0	1
1,7	4.0	2.5	2
1,11	6.0	3.0	2
2,1	1.5	4.0	3
2,2	2.0	4.5	2
2,4	3.0	5.5	2
2,7	4.5	6.0	2
2,11	6.5	6.5	2
4,1	2.5	7.0	1
4,2	3.0	7.5	2
4,4	4.0	9.0	2
4,7	5.5	11.0	1
4,11	7.5		
7,1	4.0		
7,2	4.5		
7,4	5.5		
7,7	7.0		
7,11	9.0		
11,1	6.0		
11,2	6.5		
11,4	7.5		
11,7	9.0		
11,11	11.0		

(e) $\mu_{\overline{X}} = 5.0$, $\sigma_{\overline{X}} = 2.57$ (f) $\sigma_{\overline{X}} = 2.57$ (g) Yes, 5 (h) Yes, 2.57 (i) No; however, it is more like the normal curve than is the population of 5 scores. This is in accord with the central limit theorem.

7. (a) .19 (b) .32

9. (a) 159.3 (b) 139.7–160.3

11. (a) 153.2 (b) 145.9–154.1

13. $\mu_{\overline{X}} = 100$ and $\sigma_{\overline{X}} = 1.5$ (a) .0004 (b) .0004 (c) $z = \pm 1.96$. The limits are 97.1 and 102.9. (d) Sample means show less variation as sample size increases.

CHAPTER 12

1. (a) No (b) No (c) No (d) No (e) No (f) Yes (g) No (h) No

3. If the outcome is statistically significant under a one-tailed test but not under the equivalent two-tailed test, the researcher may be tempted to report that she or he did the one-tailed test and rejected the null hypothesis. In this case, the nominal α level for the test is not the real α level, which is higher. This procedure raises the risk of wrongly rejecting the null hypothesis. In general, H_A should be determined by the logic of the inquiry, not by the data.

5. (a) False (b) False (c) False (d) False (e) False (f) True

7. No; the .95 is the probability of retaining H_0 *when* it is true, not the probability that it *is* true. There is no way to turn an α level into a number indicating the degree of confidence warranted by the conclusion about the null hypothesis.

9. $$z = \frac{87.1 - 85.0}{10/\sqrt{100}} = +2.1$$

Dr. Brown would conclude that students in her school district perform better on that particular mathematics achievement test.

11. (a) H_0: $\mu_X = 123$, H_A: $\mu_X \neq 123$ (b) $z_{\text{calc}} = -5.41$, $z_{\text{crit}} = \pm 1.96$; reject H_0 and conclude that pupils in his school district score lower on the social science test. (c) $z_{\text{calc}} = -5.41$, $z_{\text{crit}} = \pm 2.58$; reject H_0 and conclude as in part (b).

13. (a) H_0: $\mu_X = 8.3$, H_A: $\mu_X < 8.3$, $z_{\text{crit}} = -1.645$
(b) $z_{\text{calc}} = -1.68$
(c) Reject H_0
(d) With the use of a one-tailed test, we were able to reject H_0 because it took less of a difference (−1.645) from the value stated in the null hypothesis to reject the hypothesis than it did with a two-tailed test (±1.96).

15. (a) +2.602 (b) ±2.306 (c) .10 (d) $p < .01$

17. (a) $s_X = 2.45$ ($s_{\overline{X}} = 1.09$)

(b) $$t_{\text{calc}} = \frac{5.0 - 8.0}{\sqrt{\dfrac{149 - \dfrac{(25)^2}{5}}{5(5-1)}}} = -2.74$$

(c) $t_{\text{crit}} = .05$ for $df = 4$: ± 2.776, retain H_0
(d) $t_{\text{crit}} = .01$ for $df = 4$: ± 4.604, retain H_0

19. (a) H_A: $\mu_X < 98$ (b) We would have to retain the null hypothesis because the result is opposite the direction stated by the alternative hypothesis.

21. (a) $t_{\text{calc}} = +12.22$ (b) $t_{\text{crit}} = \pm 2.064$; reject H_0 and conclude that the advertisement is correct. (c) The increase in n results in a smaller value of $s_{\overline{X}}$; thus, it takes less of a difference between μ_{hyp} and μ_{obt} to reject the null hypothesis. In addition, the critical value of t is smaller.

23. (a) H_0: $\mu_X = 90$, H_A: $\mu_X < 90$, $t_{\text{crit}} = -1.761$, ($\Sigma X = 1{,}306$; $\Sigma X^2 = 114{,}404$),

$$t = \frac{87.067 - 90}{\sqrt{\dfrac{114{,}404 - \dfrac{(1306)^2}{15}}{15(14)}}}$$

$= -1.61$

Retain H_0.

25. (a) H_0: $\mu_X = 80$, H_A: $\mu_X > 80$. (b) $n = 60$, $\overline{X} = 84.68$, $s_X = 5$, $s_{\overline{X}} = 0.645$, $t_{\text{calc}} = 7.255$, $df = 59$. We reject the null hypothesis because our calculated t statistic is located at the positive end of the sampling distribution and its p-value is less than the .01 level of significance. Thus, the industrial/organizational psychologist should conclude that the new training program was effective. (Note that because IBM SPSS reports the p-value associated with the calculated t statistic for a two-tailed test, to obtain the p-value associated with the calculated t statistic for a one-tailed test the p-value should be divided by 2. Recall that the distribution of t is symmetrical.)

CHAPTER 13

1. r^2 appears small, but it is important to examine the effect size. As it turned out, the effect size for this study revealed that $r =$ .32, which meant that 66 of 100 subjects receiving psychotherapy improved, whereas only 34 of 100 subjects not receiving psychotherapy improved.

3. $\dfrac{104 - 100}{20} = .2$

5. (a) 0.64 (b) 2.44

7. (a) .92; .08 (b) .83; .17

9. (a) .83; .17 (b) .48; .52

11. (a) H_0: $\mu_X = 3.1$ H_A: $\mu_X \neq 3.1$ (b) $d = .09/.30 = .30$, so n should be slightly less than 100.

13. 70

15. (a) .094 (b) .314 (c) .656 (d) There is not a good chance of detecting a difference.

17. The authors of the study properly generalized their results only to coeds at Brown University, whereas the newspaper improperly generalized the results to all college women; nationally, college women may or may not behave similarly to those attending Brown.

CHAPTER 14

1. No; because of the variability introduced by random sampling, we would expect two samples to yield different means even if the two groups had been treated alike. We can conclude that the memory pill works only if the difference between the two sample means is so great that it could not reasonably be accounted for by chance variation when the population means are the same.

3. (a) H_0: $\mu_X - \mu_Y = 0$, H_A: $\mu_X - \mu_Y \neq 0$ (b) 1.03 (c) −9.71 (d) In both cases, reject H_0 and conclude that Y is greater than X.

5. .25; —

7. (a) H_0: $\mu_E - \mu_{NE} = 0$, H_A: $\mu_E - \mu_{NE} \neq 0$ (b) $s_{\overline{E}-\overline{NE}} = 2.91$;

$$t = \frac{(18 - 32.5) - 0}{\sqrt{\left(\dfrac{276 + 262}{15}\right)(.2361)}} = -4.98$$

Reject H_0. (c) No (d) $g = 2.42$, $r = +.79$

9. (a) H_0: $\mu_{\text{ABZ}} - \mu_{\text{placebo}} = 0$ H_A: $\mu_{\text{ABZ}} - \mu_{\text{placebo}} \neq 0$ (b) Sample (c) $s_{\overline{\text{ABZ}}-\overline{\text{P}}} = 14.8$,

$$t = \frac{(247.5 - 174.0) - 0}{\sqrt{\left(\dfrac{18{,}062.5 + 1640}{18}\right)(0.2)}} = +4.97$$

$t_{\text{crit}} = 2.878$, reject H_0 and conclude that ABZ increases the CD4+ cell count. (d) $g = 2.22$, $r = +.76$ (e) No; the researcher would have to randomly assign available subjects. For most situations in which groups are formed by random assignment, use of the random sampling model will lead to the same statistical conclusion as would result from use of the proper model.

11. (a) H_0: $\mu_X - \mu_Y = 0$ H_A: $\mu_X - \mu_Y \neq 0$ (b) $s_{\overline{X}-\overline{Y}} = 3.02$,

$$t = \frac{(40.2 - 36.2) - 0}{\sqrt{\left(\dfrac{335.6 + 487.6}{18}\right)(.2)}} = +1.32$$

$t_{\text{crit}} = 2.101$, retain H_0 (c) No; retaining H_0 means that we do not have sufficient evidence to reject H_0—it does not mean that H_0 is probably true, only that it *could* be true. (d) $g = 0.59$, $r = .30$

13. (a) H_0: $\mu_{\text{notice}} - \mu_{\text{no notice}} = 0$ H_A: $\mu_{\text{notice}} - \mu_{\text{no notice}} \neq 0$ (b) $d = .25$, n = approximately 500

15. (a) .39 × 50 replications = 19 or 20 (b) .66 × 50 = 33 (c) .96 × 50 = 48 (d) .999 × 50 = 50

17. (a) .31 (b) .95 (c) .99+ (d) Better chance of detecting a difference than in Problem 16.

19. (a) 400 (b) Approximately 66 (c) 25 (d) 13

CHAPTER 15

1. Only in that the sign of the calculated t would be different. (b) No; the difference between the two means is still the same. (c) Because subjects were tested twice (two scores for each subject).

3. (a) H_0: $\mu_X - \mu_Y = 0$, H_A: $\mu_X - \mu_Y \neq 0$ (b) $s_{\overline{X}-\overline{Y}} = 1.80$ (c) $t = +1.67$; retain H_0 at both levels

5. (a) H_0: $\mu_G - \mu_R = 0$, H_A: $\mu_G - \mu_R \neq 0$

(b) $$t_{\text{calc}} = \frac{0.6}{\sqrt{\dfrac{80 - \dfrac{(6)^2}{10}}{10(9)}}} = +0.65$$

($s_{\overline{D}} = 0.92$); retain H_0 (c) No; only the sign of t would differ. (d) $g = .21$

7. (a) H_0: $\mu_{\text{UB}} - \mu_B = 0$, H_A: $\mu_{\text{UB}} - \mu_B \neq 0$

bapp03.qxd 6/19/10 4:27 PM Page 432

(b) $t_{\text{calc}} = \dfrac{-5.3}{\sqrt{\dfrac{419 - \dfrac{(-53)^2}{10}}{10(9)}}} = -4.28$

($s_{\overline{D}} = 1.24$); reject H_0 and conclude that subjects make more accurate estimates in an unbiased environment. (c) $g = 1.35$

9. $t_{\text{calc}} = \dfrac{4.67}{\sqrt{\dfrac{978 - \dfrac{(4900)^2}{15}}{15(14)}}} = +2.65$

Reject H_0 and conclude that students' learning of arithmetic is worse under the new method.

CHAPTER 16

1. (a) .878 (b) .632 (c) .396 (d) .505 (e) .337 (f) .197 (g) .062

3. (a) $r_{\text{crit}} = \pm.374$, reject H_0 (b) $r_{\text{crit}} = \pm.590$, retain H_0 (c) $r_{\text{crit}} = -.360$, reject H_0 (d) $r_{\text{crit}} = -.337$, retain H_0

5. $r = .75$ suggests a strong relationship, but with the small sample that was used, it is not great enough to reject the null hypothesis that $\rho = 0$ at $\alpha = .05$.

7. (a) Yes; the coefficient must be $< .062$ (two-tailed test), and therefore it is probably not an important relationship. (b) No; the size of the relationship could be substantial or small enough to be unimportant.

9. (a) $df = 9$, $r_{\text{crit}} = .602$, reject H_0.

CHAPTER 17

1. (a) $63 \pm (1.984)(12/\sqrt{100})$

$C(60.62 \leq \mu_X \leq 65.38) = .95$

(b) $63 \pm (2.626)(12/\sqrt{100})$

$C(59.85 \leq \mu_X \leq 66.15) = .99$

3. (a) $117 \pm (1.990)(36/\sqrt{81})$

$C[109.04 \leq \mu_X \leq 124.96] = .95$

(b) $117 \pm (2.639)(36/\sqrt{81})$

$C[106.44 \leq \mu_X \leq 127.56] = .99$

5. **d** = .41; only moderately precise: μ_X should be within (.41 σX) of $\overline{X}$.

7. (a) $-10 \pm (1.990)\left(\sqrt{\dfrac{24^2}{36} + \dfrac{21^2}{49}}\right)$

$C[-19.95 \leq (\mu_X - \mu_Y) \leq -.05] = .95$

(b) $-10 \pm (2.639)\left(\sqrt{\dfrac{24^2}{36} + \dfrac{21^2}{49}}\right)$

$C[-23.2 \leq (\mu_X - \mu_Y) \leq +3.2] = .99$

9. $\mathbf{d} = \dfrac{(1.99)(5)}{22.5} = .44$

only moderate precision: $\mu_X - \mu_Y$ should be within .436 of $\overline{X} - \overline{Y}$.

11. $\mathbf{d} = \dfrac{(1.976)(\sqrt{8})}{17} = .33$

$\mu_X - \mu_Y$ should be within .33 of $\overline{X} - \overline{Y}$, moderately precise

13. (a) -5 (b) Reject H_0.

15. (a) $z = \dfrac{83 - 80}{15/\sqrt{625}} = +5.0$

(b) It is very unlikely that a difference this large could have arisen by chance when H_0 is true. (c) $C[81.8 \leq \mu_X \leq 84.2] = .95$ (d) Chances are high that μ_X falls within 81.8–84.2 (e) The former suggests that 80 is an unlikely value for μ_X and that 83 is far from 80; the latter suggests that the limits of sampling variation are very narrow (because of the large n), and that 80 is an unlikely value for μ_X.

17. $(247.5 - 174.0) \pm (2.878)(14.8)$

$C[30.91 \leq (\mu_X - \mu_Y) \leq 116.09] = .99$

19. $(40.2 - 36.2) \pm (2.101)(3.02)$

$C[-2.35 \leq (\mu_X - \mu_Y) \leq 10.35] = .95$

21. $0.6 \pm (2.262)(0.92)$

$C[-1.48 \leq (\mu_X - \mu_Y) \leq 2.68] = .95$

23. $-5.3 \pm (3.250)(1.24)$

$C[-9.33 \leq (\mu_X - \mu_Y) \leq -1.27] = .99$

25. 166

27. (a) $C[-.45 \leq p \leq +.76] = .95$
(b) $C[-.22 \leq p \leq +.62] = .95$
(c) $C[-.07 \leq p \leq +.52] = .95$
(d) $C[.04 \leq p \leq +.44] = .95$

CHAPTER 18

1. 1 vs. 2, 1 vs. 3, 1 vs. 4, 1 vs. 5, 1 vs. 6, 1 vs. 7, 2 vs. 3, 2 vs. 4, 2 vs. 5, 2 vs. 6, 2 vs. 7, 3 vs. 4, 3 vs. 5, 3 vs. 6, 3 vs. 7, 4 vs. 5, 4 vs. 6, 4 vs. 7, 5 vs. 6, 5 vs. 7, 6 vs. 7. With so many tests, there is an increased likelihood of committing at least one Type I error—that is, of obtaining a "significant difference" when no true difference exists.

3. Only a large value of F is evidence for rejecting H_0. When H_0 is true, $s^2_{\text{bet}}/s^2_{\text{w}}$ = unity.

When treatment effect exists, s^2_{bet} grows larger, but s^2_{w} is unaffected.

5. (a) 4, 22, 87 (b) $F_{\text{crit},.05} = 2.72$, $F_{\text{crit},.01} = 4.04$, $F_{\text{calc}} = 3.4$; difference is significant at the .05 level but not at the .01 level.

7. (a) That the mean of at least one of the groups differs from the others. (b) Make *post hoc* comparisons.

(c)

	$\bar{X}_1 =$ 14.5	$\bar{X}_2 =$ 11.5	$\bar{X}_3 =$ 19.0	$\bar{X}_4 =$ 18.1	$\bar{X}_5 =$ 12.0
$\bar{X}_1 = 14.5$	0	−3.0	4.5	3.6	−2.5
$\bar{X}_2 = 11.5$		0	7.5	6.6	0.5
$\bar{X}_3 = 19.0$			0	−0.9	3.0
$\bar{X}_4 = 18.1$				0	−6.1
$\bar{X}_5 = 12.0$					0

(d) HSD = 5.75; $\mu_2 < \mu_3$, $\mu_2 < \mu_4$, $\mu_5 < \mu_4$ (e) HSD = 7.01; $\mu_2 < \mu_3$

9. (a) H_0: $\mu_a = \mu_b = \mu_c$, (b) $SS_{\text{w}} = 18$, $SS_{\text{bet}} = 180$, $SS_{\text{total}} = 198$; $df_{\text{w}} = 11$, $df_{\text{bet}} = 2$, $df_{\text{total}} = 13$; $s^2_{\text{w}} = 1.64$, $s^2_{\text{bet}} = 90.0$, $F = 54.88$

There is a difference among the three reading programs.

(c) $\tilde{n} = 4.62$; $\text{HSD} = 3.82\sqrt{\dfrac{1.64}{4.62}} = 2.28$

(d) $\mu_{\text{B}} > \mu_{\text{A}} > \mu_{\text{C}}$

(e) $\eta^2 = 180/198 = .909$

(e) $\hat{\omega}^2_{\text{bet}} = \dfrac{180 - (2 \times 1.64)}{198 + 1.64} = .885$

11. (a) H_0: $\mu_{\text{AR}} = \mu_A = \mu_B = \mu_C$

(b)

SOURCE	*SS*	*df*	s^2	*F*
Between groups	118.0	3	39.33	5.21
Within groups	120.81	16	7.55	
Total	238.8	19		

(c) HSD = 4.98 (α = .05)

(d) $\mu_{AR} > \mu_A$, μ_B, and μ_C

(e) $\eta^2 = .49$, $\hat{\omega}^2 = .39$

13. (a) 3 (b) 1*A* vs. (1/3*B* + 1/3*C* + 1/3*D*) (c) 1*C* vs. (1/2*B* + 1/2 *D*) (d) 1*B* vs. 1*D*

15. 1 Diet vs. (1/2DW + 1/2DR), $s_k = 1.33$, $t = -6/1.33 = -4.51$

CHAPTER 19

1. c; A two-way design in which type of story and sex of subjects are the treatment variables

3. d

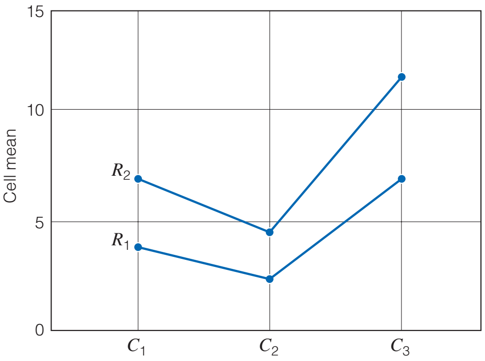
5. (a) $df_{\text{methods}} = 3$, $df_{\text{students}} = 2$, $df_{\text{wc}} = 48$, $df_{\text{methods}\times\text{students}} = 6$, $df_{\text{total}} = 59$ (b) Methods, $F_{\text{crit}} = 2.80$; students, $F_{\text{crit}} = 3.19$; interaction, $F_{\text{crit}} = 2.30$

7. (a) The interaction in the second graph is disordinal. The interactions in the other three graphs are ordinal. In the fourth graph, R_2 and R_3 cross, but the interaction is still ordinal because R_1 is greater than R_2 or R_3 at all levels of *C.* (b) In all four graphs, the main effect for rows shows a higher overall mean for row 1, yet such a conclusion does not apply at C_3 in the second graph or C_1 in the third graph (and it may not hold for C_2 in the first and third graphs and C_1 in the fourth graph).

9. (a–b)

SOURCE	*SS*	*df*	s^2	*F*	$\hat{\omega}^2$
Columns	68.17	2	34.09	21.58	.782
Rows	30.09	1	30.09	19.04	.613
Columns × rows	3.16	2	1.58	1.00	0
Within cells	9.50	6	1.58		
Total	110.92	11			

(c)



11.

Tests of Between-Subjects Effects
Dependent Variable: Performance

Source	Type III Sum of Squares	*df*	Mean Squares	*F*	Sig.
Corrected Model	101.417[a]	5	20.283	12.811	.004
Intercept	444.083	1	444.083	280.474	.000
Practice	30.083	1	30.083	19.000	.005
Motivation	68.167	2	34.083	21.526	.002
Practice *Motivation	3.167	2	1.583	1.000	.422
Error	9.500	6	1.583		
Total	555.000	12			
Corrected Total	110.917	11			

[a] R Squared = .914 (Adjusted R Squared = .843)

CHAPTER 20

1. Because all differences are squared, both positive and negative differences make a positive contribution to the value of χ^2. Only large positive values are possibly indicative of nonchance differences.

3. It would be necessary to assume that each of the 50 trials is independent of the others, which is almost certainly wrong.

5. (a) $\chi^2 = 6.93$, for $df = 2$, $\chi^2_{.05} = 5.99$, so we reject H_0; $\chi^2_{.01} = 9.21$, so we retain H_0. When χ^2 is significant, we can conclude only that H_0 is untrue in some (any) way. (b) .29 (c) Medium

7. (a) H_0: $P_A = P_B = P_C = P_D = P_E$ (b) $\chi^2 = 38.9$; for $df = 4$, $\chi^2_{.01} = 13.28$, reject H_0. (c) Conclude that the students were not guessing (i.e., responses were not random). (d) .62; large

9. (a) $\chi^2 = 127.14$, reject H_0 and conclude that more students are "helped" by changing answers. (b) .76

11. (a) Attitude about the proposition is independent of the sex of the respondent. (b) $\chi^2 = 3.13$ ($\chi^2_{.05} = 5.99$); retain H_0 (c) No interpretation because H_0 was not rejected.

13. (a) H_0: $P_A - P_B = 0$, H_A: $P_A - P_B \neq 0$ (b) $\chi^2 = 2.12$ (c) Retain H_0 (d) $\phi = .21$, $OR = \frac{(7 \times 13)}{(18 \times 12)} = 1.023$

15. (a) Attitudes toward the bond issue are independent of political affiliation. (b) $\chi^2 = 97.62$ ($df = 4$, $\chi^2_{.05} = 9.49$) (c) Reject H_0 (d) $\phi = .40$, $\hat{\omega} = .57$ (e) Because the observations are independent of one another.

CHAPTER 21

1. Parametric tests evaluate hypotheses about population parameters, whereas nonparametric tests do not. Distribution-free tests do not make any assumptions about the shape of the population. The terms *nonparametric test* and *distribution-free test* do not mean the same thing. Assumption-freer tests do not make the usual assumption about population distributions, but the tests are not completely free of assumptions.

3. The sampling distribution of a randomization test is derived empirically, whereas the sampling distribution of *t* is derived through theory.

5. Converting original scores to ranks effectively eliminates outliers (scores that are very different from others in the distribution). For example, when ranked, the scores 6, 8, 9, 10, 100 become 1, 2, 3, 4, 5. The presence of an outlier can often eliminate the possibility of obtaining a significant result with a parametric test.

7. No; $\Sigma R_X = 45$, which is within the critical values 21–49 for the Mann-Whitney *U* test. $Mdn_{\text{quiet}} = 59$, $Mdn_{\text{noisy}} = 51$; those in the quiet environment performed better, but the difference was not significant.

9. $\Sigma R_X = 30.5$; critical values for a two-tailed test at the .01 level are 34–78, so the null hypothesis is rejected. Androgyny appears to be predictive of longer-lasting relationships.

11. $\chi^2_{\text{calc}} = 4.08$, $df = 1$; $\chi^2_{.05} = 3.84$, reject H_0; $\chi^2_{.01} = 6.63$, retain H_0.

13. (a) $R_- = 9$, $R_{\text{crit}} = 10$, reject H_0. (b) $r_c = .73$ (c) No; factors other than the film may have been operative; it is not clear that the film was responsible for the difference.

15. (a) $R_- = 3$, $R_{\text{crit}} = 3$, reject H_0. (b) $r_c = .83$

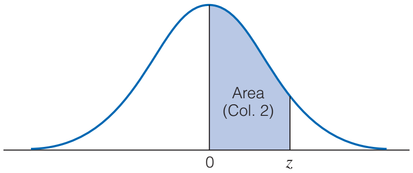
17. $\chi^2_F = 6.5$, at $\alpha = .05$ and 2 *df*, reject H_0; $W = .36$

APPENDIX D

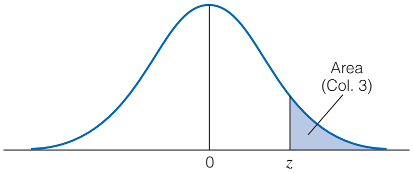
Statistical Tables

TABLE A *Areas under the Normal Curve Corresponding to Given Values of z*

Column 2 gives the proportion of the area under the entire curve that is between the mean ($z = 0$) and the positive value of z. Areas for negative values of z are the same as for positive values because the curve is symmetrical.



Column 3 gives the proportion of the area under the entire curve that falls beyond the stated positive value of z. Areas for negative values of z are the same because the curve is symmetrical.



z 1	AREA BETWEEN MEAN AND *z* 2	AREA BEYOND *z* 3	*z* 1	AREA BETWEEN MEAN AND *z* 2	AREA BEYOND *z* 3
0.00	.0000	.5000	0.27	.1064	.3936
0.01	.0040	.4960	0.28	.1103	.3897
0.02	.0080	.4920	0.29	.1141	.3859
0.03	.0120	.4880	0.30	.1179	.3821
0.04	.0160	.4840	0.31	.1217	.3783
0.05	.0199	.4801	0.32	.1255	.3745
0.06	.0239	.4761	0.33	.1293	.3707
0.07	.0279	.4721	0.34	.1331	.3669
0.08	.0319	.4681	0.35	.1368	.3632
0.09	.0359	.4641	0.36	.1406	.3594
0.10	.0398	.4602	0.37	.1443	.3557
0.11	.0438	.4562	0.38	.1480	.3520
0.12	.0478	.4522	0.39	.1517	.3483
0.13	.0517	.4483	0.40	.1554	.3446
0.14	.0557	.4443	0.41	.1591	.3409
0.15	.0596	.4404	0.42	.1628	.3372
0.16	.0636	.4364	0.43	.1664	.3336
0.17	.0675	.4325	0.44	.1700	.3300
0.18	.0714	.4286	0.45	.1736	.3264
0.19	.0753	.4247	0.46	.1772	.3228
0.20	.0793	.4207	0.47	.1808	.3192
0.21	.0832	.4168	0.48	.1844	.3156
0.22	.0871	.4129	0.49	.1879	.3121
0.23	.0910	.4090	0.50	.1915	.3085
0.24	.0948	.4052	0.51	.1950	.3050
0.25	.0987	.4013	0.52	.1985	.3015
0.26	.1026	.3974	0.53	.2019	.2981
0.54	.2054	.2946	0.92	.3212	.1788
0.55	.2088	.2912	0.93	.3238	.1762
0.56	.2123	.2877	0.94	.3264	.1736
0.57	.2157	.2843	0.97	.3340	.1660
0.58	.2190	.2810	0.98	.3365	.1635
0.59	.2224	.2776	0.99	.3389	.1611

TABLE A (*Continued*)

z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3	z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3
0.60	.2257	.2743	1.00	.3413	.1587
0.61	.2291	.2709	1.01	.3438	.1562
0.62	.2324	.2676	1.02	.3461	.1539
0.63	.2357	.2643	1.03	.3485	.1515
0.64	.2389	.2611	1.04	.3508	.1492
0.65	.2422	.2578	1.05	.3531	.1469
0.66	.2454	.2546	1.06	.3554	.1446
0.67	.2486	.2514	1.07	.3577	.1423
0.68	.2517	.2483	1.08	.3599	.1401
0.69	.2549	.2451	1.09	.3621	.1379
0.70	.2580	.2420	1.10	.3643	.1357
0.71	.2611	.2389	1.11	.3665	.1335
0.72	.2642	.2358	1.12	.3686	.1314
0.73	.2673	.2327	1.13	.3708	.1292
0.74	.2704	.2296	1.14	.3729	.1271
0.75	.2734	.2266	1.15	.3749	.1251
0.76	.2764	.2236	1.16	.3770	.1230
0.77	.2794	.2206	1.17	.3790	.1210
0.78	.2823	.2177	1.18	.3810	.1190
0.79	.2852	.2148	1.19	.3830	.1170
0.80	.2881	.2119	1.20	.3849	.1151
0.81	.2910	.2090	1.21	.3869	.1131
0.82	.2939	.2061	1.22	.3888	.1112
0.83	.2967	.2033	1.23	.3907	.1093
0.84	.2995	.2005	1.24	.3925	.1075
0.85	.3023	.1977	1.25	.3944	.1056
0.86	.3051	.1949	1.26	.3962	.1038
0.87	.3078	.1922	1.27	.3980	.1020
0.88	.3106	.1894	1.28	.3997	.1003
0.89	.3133	.1867	1.29	.4015	.0985
0.90	.3159	.1841	1.30	.4032	.0968
0.91	.3186	.1814	1.31	.4049	.0951
1.32	.4066	.0934	1.70	.4554	.0446
1.33	.4082	.0918	1.71	.4564	.0436
1.34	.4099	.0901	1.72	.4573	.0427
1.35	.4115	.0885	1.73	.4582	.0418
1.36	.4131	.0869	1.74	.4591	.0409
1.37	.4147	.0853	1.75	.4599	.0401
1.38	.4162	.0838	1.76	.4608	.0392
1.39	.4177	.0823	1.77	.4616	.0384
1.40	.4192	.0808	1.78	.4625	.0375
1.41	.4207	.0793	1.79	.4633	.0367
1.42	.4222	.0778	1.80	.4641	.0359
1.43	.4236	.0764	1.81	.4649	.0351
1.44	.4251	.0749	1.82	.4656	.0344

(*continued*)

TABLE A *(Continued)*

z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3	z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3
1.45	.4265	.0735	1.83	.4664	.0336
1.46	.4279	.0721	1.84	.4671	.0329
1.47	.4292	.0708	1.85	.4678	.0322
1.48	.4306	.0694	1.86	.4686	.0314
1.49	.4319	.0681	1.87	.4693	.0307
1.50	.4332	.0668	1.88	.4699	.0301
1.51	.4345	.0655	1.89	.4706	.0294
1.52	.4357	.0643	1.90	.4713	.0287
1.53	.4370	.0630	1.91	.4719	.0281
1.54	.4382	.0618	1.92	.4726	.0274
1.55	.4394	.0606	1.93	.4732	.0268
1.56	.4406	.0594	1.94	.4738	.0262
1.57	.4418	.0582	1.95	.4744	.0256
1.58	.4429	.0571	1.96	.4750	.0250
1.59	.4441	.0559	1.97	.4756	.0244
1.60	.4452	.0548	1.98	.4761	.0239
1.61	.4463	.0537	1.99	.4767	.0233
1.62	.4474	.0526	2.00	.4772	.0228
1.63	.4484	.0516	2.01	.4778	.0222
1.64	.4495	.0505	2.02	.4783	.0217
1.65	.4505	.0495	2.03	.4788	.0212
1.66	.4515	.0485	2.04	.4793	.0207
1.67	.4525	.0475	2.05	.4798	.0202
1.68	.4535	.0465	2.06	.4803	.0197
1.69	.4545	.0455	2.07	.4808	.0192
2.08	.4812	.0188	2.43	.4925	.0075
2.09	.4817	.0183	2.44	.4927	.0073
2.10	.4821	.0179	2.45	.4929	.0071
2.11	.4826	.0174	2.46	.4931	.0069
2.12	.4830	.0170	2.47	.4932	.0068
2.13	.4834	.0166	2.48	.4934	.0066
2.14	.4838	.0162	2.49	.4936	.0064
2.15	.4842	.0158	2.50	.4938	.0062
2.16	.4846	.0154	2.51	.4940	.0060
2.17	.4850	.0150	2.52	.4941	.0059
2.18	.4854	.0146	2.53	.4943	.0057
2.19	.4857	.0143	2.54	.4945	.0055
2.20	.4861	.0139	2.55	.4946	.0054
2.21	.4864	.0136	2.56	.4948	.0052
2.22	.4868	.0132	2.57	.4949	.0051
2.23	.4871	.0129	2.58	.4951	.0049
2.24	.4875	.0125	2.59	.4952	.0048
2.25	.4878	.0122	2.60	.4953	.0047
2.26	.4881	.0119	2.61	.4955	.0045
2.27	.4884	.0116	2.62	.4956	.0044

TABLE A (*Continued*)

z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3	z 1	AREA BETWEEN MEAN AND z 2	AREA BEYOND z 3
2.28	.4887	.0113	2.63	.4957	.0043
2.29	.4890	.0110	2.64	.4959	.0041
2.30	.4893	.0107	2.65	.4960	.0040
2.31	.4896	.0104	2.66	.4961	.0039
2.32	.4898	.0102	2.67	.4962	.0038
2.33	.4901	.0099	2.68	.4963	.0037
2.34	.4904	.0096	2.69	.4964	.0036
2.35	.4906	.0094	2.70	.4965	.0035
2.36	.4909	.0091	2.71	.4966	.0034
2.37	.4911	.0089	2.72	.4967	.0033
2.38	.4913	.0087	2.73	.4968	.0032
2.39	.4916	.0084	2.74	.4969	.0031
2.40	.4918	.0082	2.75	.4970	.0030
2.41	.4920	.0080	2.76	.4971	.0029
2.42	.4922	.0078	2.77	.4972	.0028
2.78	.4973	.0027	3.04	.4988	.0012
2.79	.4974	.0026	3.05	.4989	.0011
2.80	.4974	.0026	3.06	.4989	.0011
2.81	.4975	.0025	3.07	.4989	.0011
2.82	.4976	.0024	3.08	.4990	.0010
2.83	.4977	.0023	3.09	.4990	.0010
2.84	.4977	.0023	3.10	.4990	.0010
2.85	.4978	.0022	3.11	.4991	.0009
2.86	.4979	.0021	3.12	.4991	.0009
2.87	.4979	.0021	3.13	.4991	.0009
2.88	.4980	.0020	3.14	.4992	.0008
2.89	.4981	.0019	3.15	.4992	.0008
2.90	.4981	.0019	3.16	.4992	.0008
2.91	.4982	.0018	3.17	.4992	.0008
2.92	.4982	.0018	3.18	.4993	.0007
2.93	.4983	.0017	3.19	.4993	.0007
2.94	.4984	.0016	3.20	.4993	.0007
2.95	.4984	.0016	3.21	.4993	.0007
2.96	.4985	.0015	3.22	.4994	.0006
2.97	.4985	.0015	3.23	.4994	.0006
2.98	.4986	.0014	3.24	.4994	.0006
2.99	.4986	.0014	3.30	.4995	.0005
3.00	.4987	.0013	3.40	.4997	.0003
3.01	.4987	.0013	3.50	.4998	.0002
3.02	.4987	.0013	3.60	.4998	.0002
3.03	.4988	.0012	3.70	.4999	.0001

Source: From Appendix 2 of R. Clarke, A. Coladarci, and J. Caffrey, *Statistical Reasoning and Procedures*, Charles E. Merrill Books, with permission of the author.

TABLE B *The Binomial Distribution*

N	NUMBER OF *P* OR *Q* EVENTS	*P* or *Q*									
		.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0277	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641

TABLE B (*Continued*)

N	NUMBER OF *P* OR *Q* EVENTS	*P* or *Q*									
		.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0005
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0054
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1231	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095

(*continued*)

TABLE B (*Continued*)

		P or *Q*									
N	NUMBER OF *P* OR *Q* EVENTS	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Source: Reprinted from *Handbook of Probability and Statistics with Tables* by R. S. Burington and D. C. May, 2d ed. McGraw-Hill Book Co., 1970, by permission of the author's trustee.

TABLE C *Random Numbers*

11339	19233	50911	14209	39594	68368	97742	36252	27671	55091
96971	19968	31709	40197	16313	80020	01588	21654	50328	04577
07779	47712	33846	84716	49870	59670	46946	71716	50623	38681
71675	95993	08790	13241	71260	16558	83316	68482	10294	45137
32804	72742	16237	72550	10570	31470	92612	94917	48822	79794
14835	56263	53062	71543	67632	30337	28739	17582	40924	32434
15544	14327	07580	48813	30161	10746	96470	60680	63507	14435
92230	41243	90765	08867	08038	05038	10908	00633	21740	55450
33564	93563	10770	10595	71323	84243	09402	62877	49762	56151
84461	55618	40570	72906	30794	49144	65239	21788	38288	29180
91645	42451	83776	99246	45548	02457	74804	49536	89815	74285
78305	63797	26995	23146	56071	97081	22376	09819	56855	97424
97888	55122	65545	02904	40042	70653	24483	31258	96475	77668
67286	09001	09718	67231	54033	24185	52097	78713	95910	84400
53610	59459	89945	72102	66595	02198	26968	88467	46939	52318
52965	76189	68892	64541	02225	09603	59304	38179	75920	80486
25336	39735	25594	50557	96257	59700	27715	42432	27652	88151
73078	44371	77616	49296	55882	71507	30168	31876	28283	53424
31797	52244	38354	47800	48454	43304	14256	74281	82279	28882
47772	22798	36910	39986	34033	39868	24009	97123	59151	27583
54153	70832	37575	31898	39212	63993	05419	77565	73150	98537
93745	99871	37129	55032	94444	17884	27082	23502	06136	89476
81676	51330	58828	74199	87214	13727	80539	95037	73536	16862
79788	02193	33250	05865	53018	62394	56997	41534	01953	13763
92112	61235	68760	61201	02189	09424	24156	10368	26527	89107
87542	28171	45150	75523	66790	63963	13903	68498	02981	25219
37535	48342	48943	07719	20407	33748	93650	39356	01011	22099
95957	96668	69380	49091	90182	13205	71802	35482	27973	46814
34642	85350	53361	63940	79546	89956	96836	81313	80712	73572
50413	31008	09231	46516	61672	79954	01291	72278	55658	84893
53312	73768	59931	55182	43761	59424	79775	17772	41552	45236
16302	64092	76045	28958	21182	30050	96256	85737	86962	27067
96357	98654	01909	58799	87374	53184	87233	55275	59572	56476
38529	89095	89538	15600	33687	86353	61917	63876	52367	79032
45939	05014	06099	76041	57638	55342	41269	96173	94872	35605
02300	23739	68485	98567	77035	91533	62500	31548	09511	80252
59750	14131	24973	05962	83215	25950	43867	75213	21500	17758
21285	53607	82657	22053	29996	04729	48917	72091	57336	18476
93703	60164	19090	63030	88931	84439	94747	77982	61932	21928
15576	76654	19775	77518	43259	82790	08193	63007	68824	75315
12752	33321	69796	03625	37328	75200	77262	99004	96705	15540
89038	53455	93322	25069	88186	45026	31020	52540	10838	72490
62411	56968	08379	40159	27419	12024	99694	68668	73039	87682
45853	68103	38927	77105	65241	70387	01634	59665	30512	66161
84558	24272	84355	00116	68344	92805	52618	51584	75964	53021
45272	58388	69131	61075	80192	45959	76992	19210	27126	45525
68015	99001	11832	39832	80462	70468	89929	55695	77524	20675
13263	92240	89559	66545	06433	38634	36645	22350	81169	97417
66309	31466	97705	46996	69059	33771	95004	89037	38054	80853
56348	05291	38713	82303	26293	61319	45285	75784	50043	44438

(continued)

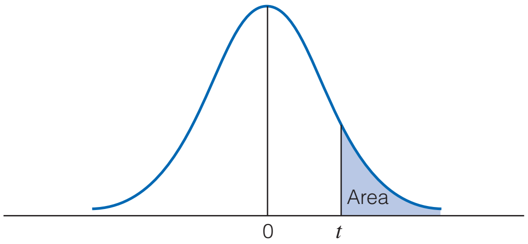
TABLE C *(Continued)*

93108	77033	68325	10160	38667	62441	87023	94372	06164	30700
28271	08589	83279	48838	60935	70541	53814	95588	05832	80235
21841	35545	11148	34775	17308	88034	97765	35959	52843	44895
22025	79554	19698	25255	50283	94037	57463	92925	12042	91414
09210	20779	02994	02258	86978	85092	54052	18354	20914	28460
90552	71129	03621	20517	16908	06668	29916	51537	93658	29525
01130	06995	20258	10351	99248	51660	38861	49668	74742	47181
22604	56719	21784	68788	38358	59827	19270	99287	81193	43366
06690	01800	34272	65497	94891	14537	91358	21587	95765	72605
59809	69982	71809	64984	48709	43991	24987	69246	86400	29559
56475	02726	58511	95405	70293	84971	06676	44075	32338	31980
02730	34870	83209	03138	07715	31557	55242	61308	26507	06186
74482	33990	13509	92588	10462	76546	46097	01825	20153	36271
19793	22487	94238	81054	95488	23617	15539	94335	73822	93481
19020	27856	60526	24144	98021	60564	46373	86928	52135	74919
69565	60635	65709	77887	42766	86698	14004	94577	27936	47220
69274	23208	61035	84263	15034	28717	76146	22021	23779	98562
83658	14204	09445	41081	49630	34215	89806	40930	97194	21747
78612	51102	66826	40430	54072	62164	68977	95583	11765	81072
14980	74158	78216	38985	60838	82836	42777	85321	90463	11813
63172	28010	29405	91554	75195	51183	65805	87525	35952	83204
71167	37984	52737	06869	38122	95322	41356	19391	96787	64410
78530	56410	19195	34434	83712	50397	80920	15464	81350	18673
98324	03774	07573	67864	06497	20758	83454	22756	83959	96347
55793	30055	08373	32652	02654	75980	02095	87545	88815	80086
05674	34471	61967	91266	38814	44728	32455	17057	08339	93997
15643	22245	07592	22078	73628	60902	41561	54608	41023	98345
66750	19609	70358	03622	64898	82220	69304	46235	97332	64539
42320	74314	50222	82339	51564	42885	50482	98501	02245	88990
73752	73818	15470	04914	24936	65514	56633	72030	30856	85183
97546	02188	46373	21486	28221	08155	23486	66134	88799	49496
32569	52162	38444	42004	78011	16909	94194	79732	47114	23919
36048	93973	82596	28739	86985	58144	65007	08786	14826	04896
40455	36702	38965	56042	80023	28169	04174	65533	52718	55255
33597	47071	55618	51796	71027	46690	08002	45066	02870	60012
22828	96380	35883	15910	17211	42358	14056	55438	98148	35384
00631	95925	19324	31497	88118	06283	84596	72091	53987	01477
75722	36478	07634	63114	27164	15467	03983	09141	60562	65725
80577	01771	61510	17099	28731	41426	18853	41523	14914	76661
10524	20900	65463	83680	05005	11611	64426	59065	06758	02892
93815	69446	75253	51915	97839	75427	90685	60352	96288	34248
81867	97119	93446	20862	46591	97677	42704	13718	44975	67145
64649	07689	16711	12169	15238	74106	60655	56289	74166	78561
55768	09210	52439	33355	57884	36791	00853	49969	74814	09270
38080	49460	48137	61589	42742	92035	21766	19435	92579	27683
22360	16332	05343	34613	24013	98831	17157	44089	07366	66196
40521	09057	00239	51284	71556	22605	41293	54854	39736	05113
19292	69862	59951	49644	53486	28244	20714	56030	39292	45166
79504	40078	06838	05509	68581	39400	85615	52314	83202	40313
64138	27983	84048	42631	58658	62243	82572	45211	37060	15017

Source: Reprinted from *A Million Random Digits with 100,000 Normal Deviates*, RAND (New York: The Free Press, 1955). Copyright © 1955 and 1983 by RAND. Used by permission.

TABLE D *Student's* t *Distribution*

The first column identifies the specific *t* distribution according to its number of degrees of freedom. Other columns give the proportion of the area under the entire curve that falls beyond the tabled positive value of *t*. Areas for negative values of *t* are the same because the curve is symmetrical.



	AREA IN ONE TAIL					
df	**.25**	**.10**	**.05**	**.025**	**.01**	**.005**
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.500
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.696	1.356	1.782	2.179	2.681	3.054
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.132	2.602	2.947
16	0.690	1.337	1.746	2.120	2.584	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.540	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.568	2.819
23	0.685	1.320	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.312	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
31	0.682	1.310	1.696	2.040	2.453	2.744
32	0.682	1.309	1.694	2.037	2.449	2.738
33	0.682	1.308	1.692	2.034	2.445	2.733
34	0.682	1.307	1.691	2.032	2.441	2.728
35	0.682	1.306	1.690	2.030	2.438	2.724
36	0.681	1.306	1.688	2.028	2.434	2.720
37	0.681	1.305	1.687	2.026	2.431	2.715
38	0.681	1.304	1.686	2.024	2.429	2.712
39	0.681	1.304	1.685	2.023	2.426	2.708
40	0.681	1.303	1.684	2.021	2.423	2.704
45	0.680	1.301	1.679	2.014	2.412	2.690

(continued)

TABLE D (*Continued*)

	AREA IN ONE TAIL					
df	.25	.10	.05	.025	.01	.005
50	0.679	1.299	1.676	2.009	2.403	2.678
55	0.679	1.297	1.673	2.004	2.396	2.668
60	0.679	1.296	1.671	2.000	2.390	2.660
70	0.678	1.294	1.667	1.994	2.381	2.648
80	0.678	1.292	1.664	1.990	2.374	2.639
90	0.677	1.291	1.662	1.987	2.368	2.632
100	0.677	1.290	1.660	1.984	2.364	2.626
120	0.676	1.289	1.658	1.980	2.358	2.617
150	0.676	1.287	1.655	1.976	2.352	2.609
200	0.676	1.286	1.652	1.972	2.345	2.601
300	0.675	1.284	1.650	1.968	2.339	2.592
400	0.675	1.284	1.649	1.966	2.336	2.588
500	0.675	1.283	1.648	1.965	2.334	2.586
1000	0.675	1.282	1.646	1.962	2.330	2.581
∞	0.674	1.282	1.645	1.960	2.326	2.576

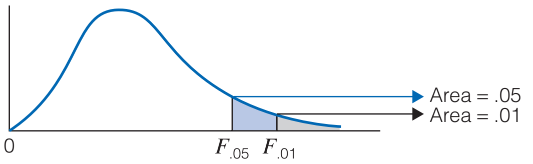
Adapted from *Handbook of Statistical Tables* by Owen, p. 28–30. © 1962 by Addison-Wesley Publishing Company, Inc. Copyright renewal © 1990. Reprinted by permission of Pearson Education, Inc. All rights reserved.

LibraryPirate

bapp04.qxd 6/16/10 12:34 PM Page 447

TABLE E *The F Distribution*

The specific F distribution must be identified by the number of degrees of freedom characterizing the numerator and the denominator of F. The values of F corresponding to 5% of the area in the upper tail are shown in lightface type. Those corresponding to 1% are in boldface type.



DEGREES OF FREEDOM: DENOMINATOR	DEGREES OF FREEDOM: NUMERATOR																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	241	252	253	253	254	254	254
	4052	**4999**	**5403**	**5625**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6082**	**6106**	**6142**	**6169**	**6208**	**6234**	**6258**	**6286**	**6302**	**6323**	**6334**	**6352**	**6361**	**6366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	3.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**

(continued)

LibraryPirate

TABLE E (*Continued*)

DEGREES OF FREEDOM: DENOMINATOR	DEGREES OF FREEDOM: NUMERATOR																							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**14**	**16**	**20**	**24**	**30**	**40**	**50**	**75**	**100**	**200**	**500**	**∞**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	2.98	1.95	1.93	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**

LibraryPirate

TABLE E *(Continued)*

DEGREES OF FREEDOM: DENOMINATOR	DEGREES OF FREEDOM: NUMERATOR																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.63
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**
46	4.04	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**

(continued)

LibraryPirate

bapp04.qxd 6/16/10 12:34 PM Page 450

TABLE E *(Continued)*

DEGREES OF FREEDOM: DENOMINATOR	DEGREES OF FREEDOM: NUMERATOR																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**

Source: Reproduced by permission from *Statistical Methods*, 5th ed., by George W. Snedecor. Copyright © 1956 by Iowa State University Press.

TABLE F *The Studentized Range Statistic*

		k **= NUMBER OF GROUPS**								
df_W	α	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16

Source: Abridged from Table 29, E. Pearson and H. Hartley, *Biometrika Tables for Statisticians*, vol. 1, 3d ed., University Press, Cambridge, 1966, with permission of the Biometrika Trustees.

TABLE G *Values of the Correlation Coefficient Required for Different Levels of Significance When* $H_0: \rho = 0$

	LEVELS OF SIGNIFICANCE FOR A ONE-TAILED TEST			
	.05	**.025**	**.01**	**.005**
	LEVELS OF SIGNIFICANCE FOR A TWO-TAILED TEST			
df	**.10**	**.05**	**.02**	**.01**
1	.988	.977	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
32	.287	.339	.397	.436
34	.279	.329	.386	.424
36	.271	.320	.376	.413
38	.264	.312	.367	.403
40	.257	.304	.358	.393
42	.251	.297	.350	.384
44	.426	.291	.342	.376
46	.240	.285	.335	.368
48	.235	.279	.328	.361
50	.231	.273	.322	.354
55	.220	.261	.307	.339
60	.211	.250	.295	.325
65	.203	.240	.284	.313
70	.195	.232	.274	.302
75	.189	.224	.265	.292

TABLE G (*Continued*)

	LEVELS OF SIGNIFICANCE FOR A ONE-TAILED TEST			
	.05	.025	.01	.005
	LEVELS OF SIGNIFICANCE FOR A TWO-TAILED TEST			
df	.10	.05	.02	.01
80	.183	.217	.256	.283
85	.178	.211	.249	.275
90	.173	.205	.242	.267
95	.168	.200	.236	.260
100	.164	.195	.230	.254
120	.150	.178	.210	.232
150	.134	.159	.189	.208
200	.116	.138	.164	.181
300	.095	.113	.134	.148
400	.082	.098	.116	.128
500	.073	.088	.104	.115
1000	.052	.062	.073	.081

Source: Table V.A of Fisher, *Statistical Methods for Research Workers*, by permission of Oxford University Press. Supplementary values were calculated at San Jose State University by K. Fernandes.

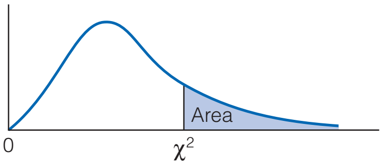
TABLE H *Values of Fisher's* z′ *for Values of* r

r	*z′*	*r*	*z′*	*r*	*z′*	*r*	*z′*	*r*	*z′*
.000	.000	.200	.203	.400	.424	.600	.693	.800	1.099
.005	.005	.205	.208	.405	.430	.605	.701	.805	1.113
.010	.010	.210	.213	.410	.436	.610	.709	.810	1.127
.015	.015	.215	.218	.415	.442	.615	.717	.815	1.142
.020	.020	.220	.224	.420	.448	.620	.725	.820	1.157
.025	.025	.225	.229	.425	.454	.625	.733	.825	1.172
.030	.030	.230	.234	.430	.460	.630	.741	.830	1.188
.035	.035	.235	.239	.435	.466	.635	.750	.835	1.204
.040	.040	.240	.245	.440	.472	.640	.758	.840	1.221
.045	.045	.245	.250	.445	.478	.645	.767	.845	1.238
.050	.050	.250	.255	.450	.485	.650	.775	.850	1.256
.055	.055	.255	.261	.455	.491	.655	.784	.855	1.274
.060	.060	.260	.266	.460	.497	.660	.793	.860	1.293
.065	.065	.265	.271	.465	.504	.665	.802	.865	1.313
.070	.070	.270	.277	.470	.510	.670	.811	.870	1.333
.075	.075	.275	.282	.475	.517	.675	.820	.875	1.354
.080	.080	.280	.288	.480	.523	.680	.829	.880	1.376
.085	.085	.285	.293	.485	.530	.685	.838	.885	1.398
.090	.090	.290	.299	.490	.536	.690	.848	.890	1.422
.095	.095	.295	.304	.495	.543	.695	.858	.895	1.447
.100	.100	.300	.310	.500	.549	.700	.867	.900	1.472
.105	.105	.305	.315	.505	.556	.705	.877	.905	1.499
.110	.110	.310	.321	.510	.563	.710	.887	.910	1.528
.115	.116	.315	.326	.515	.570	.715	.897	.915	1.557
.120	.121	.320	.332	.520	.576	.720	.908	.920	1.589
.125	.126	.325	.337	.525	.583	.725	.918	.925	1.623
.130	.131	.330	.343	.530	.590	.730	.929	.930	1.658
.135	.136	.335	.348	.535	.597	.735	.940	.935	1.697
.140	.141	.340	.354	.540	.604	.740	.950	.940	1.738
.145	.146	.345	.360	.545	.611	.745	.962	.945	1.783
.150	.151	.350	.365	.550	.618	.750	.973	.950	1.832
.155	.156	.355	.371	.555	.626	.755	.984	.955	1.886
.160	.161	.360	.377	.560	.633	.760	.996	.960	1.946
.165	.167	.365	.383	.565	.640	.765	1.008	.965	2.014
.170	.172	.370	.388	.570	.648	.770	1.020	.970	2.092
.175	.177	.375	.394	.575	.655	.775	1.033	.975	2.185
.180	.182	.380	.400	.580	.662	.780	1.045	.980	2.298
.185	.187	.385	.406	.585	.670	.785	1.058	.985	2.443
.190	.192	.390	.412	.590	.678	.790	1.071	.990	2.647
.195	.198	.395	.418	.595	.685	.795	1.085	.995	2.994

Source: From *Statistical Methods*, 2d ed., by Allen L. Edwards. Copyright © 1954, 1967 by Allen L. Edwards. Reprinted by permission of Allen L. Edwards.

TABLE I *The χ^2 Distribution*

The first column identifies the specific χ^2 distribution according to its number of degrees of freedom.* Other columns give the proportion of the area under the entire curve that falls above the tabled value of χ^2.



	AREA IN THE UPPER TAIL									
df	**.995**	**.99**	**.975**	**.95**	**.90**	**.10**	**.05**	**.025**	**.01**	**.005**
1	.000039	.00016	.00098	.0039	.016	2.71	3.84	5.02	6.63	7.88
2	.010	.020	.051	.10	.21	4.61	5.99	7.38	9.21	10.60
3	.072	.11	.22	.35	.58	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	65.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

---

*When $df > 30$, the critical value of χ^2 may be found by the following approximate formula: $\chi^2 = df[1 - (2/9df) + z\sqrt{2/9\ df}\,]^3$, where z is the normal deviate above which lies the same proportionate area in the normal curve. For example, to find the value of χ^2 that divides the upper 1% of the distribution from the remainder when $df = 30$, we calculate: $\chi^2 = 30(1 - .00741 + 2.3263\sqrt{.0074074})^3 = 50.91$, which compares closely with the tabled value of 50.89.

Source: Modified from Table 8, E. Pearson and H. Hartley, *Biometrika Tables for Statisticians*, vol. 1, 3d ed., University Press, Cambridge, 1966, with permission of the Biometrika Trustees.

TABLE J *Critical One-Tail Values of* $\sum R_X$ *for the Mann-Whitney* U *Test*

$n_X = 3$

n_Y	.005	.01	.025	.05
3				6–15
4				6–18
5			6–21	7–20
6			7–23	8–22
7		6–27	7–26	8–25
8		6–30	8–28	9–27
9	6–33	7–32	8–31	10–29
10	6–36	7–35	9–33	10–32
11	6–39	7–38	9–36	11–34
12	7–41	8–40	10–38	11–37
13	7–44	8–43	10–41	12–39
14	7–47	8–46	11–43	13–41
15	8–49	9–48	11–46	13–44

$n_X = 4$

n_Y	.005	.01	.025	.05
4			10–26	11–25
5		10–30	11–29	12–28
6	10–34	11–33	12–32	13–31
7	10–38	11–37	13–35	14–34
8	11–41	12–40	14–38	15–37
9	11–45	13–43	14–42	16–40
10	12–48	13–47	15–45	17–43
11	12–52	14–50	16–48	18–46
12	13–55	15–53	17–51	19–49
13	13–59	15–57	18–54	20–52
14	14–62	16–60	19–57	21–55
15	15–65	17–63	20–60	22–58

$n_X = 5$

n_Y	.005	.01	.025	.05
5	15–40	16–39	17–38	19–36
6	16–44	17–43	18–42	20–40
7	16–49	18–47	20–45	21–44
8	17–53	19–51	21–49	23–47
9	18–57	20–55	22–53	24–51
10	19–61	21–59	23–57	26–54
11	20–65	22–63	24–61	27–58
12	21–69	23–67	26–64	28–62
13	22–73	24–71	27–68	30–65
14	22–78	25–75	28–72	31–69
15	23–82	26–79	29–76	33–72

$n_X = 6$

n_Y	.005	.01	.025	.05
6	23–55	24–54	26–52	28–50
7	24–60	25–59	27–57	29–55
8	25–65	27–63	29–61	31–59
9	26–70	28–68	31–65	33–63
10	27–75	29–73	32–70	35–67
11	29–80	30–78	34–74	37–71
12	30–84	32–82	35–79	38–76
13	31–89	33–87	37–83	40–80
14	32–94	34–92	38–88	42–84
15	33–99	36–96	40–92	44–88

$n_X = 7$

n_Y	.005	.01	.025	.05
7	32–73	34–71	36–69	39–66
8	34–78	35–77	38–74	41–71
9	34–84	37–82	40–79	43–76
10	37–89	39–87	42–84	45–81
11	38–95	40–93	44–89	47–86
12	40–100	42–98	46–94	49–91
13	41–106	44–103	48–99	52–95
14	43–111	45–109	50–104	54–100
15	44–117	47–114	52–109	56–105

$n_X = 8$

n_Y	.005	.01	.025	.05
8	43–93	45–91	49–87	51–85
9	45–99	47–97	51–93	54–90
10	47–105	49–103	53–99	56–96
11	49–111	51–109	55–105	59–101
12	51–117	53–115	58–110	62–106
13	53–123	56–120	60–116	64–112
14	54–130	58–126	62–122	67–117
15	56–136	60–132	65–127	69–123

TABLE J (*Continued*)

$n_X = 9$

n_Y	.005	.01	.025	.05
9	56–115	59–112	62–109	66–105
10	58–122	61–119	65–115	69–111
11	61–128	63–126	68–121	72–117
12	63–135	66–132	71–127	75–123
13	65–142	68–139	73–134	78–129
14	67–149	71–145	76–140	81–135
15	69–156	73–152	79–146	84–141

$n_X = 10$

n_Y	.005	.01	.025	.05
10	71–139	74–136	78–132	82–128
11	73–147	77–143	81–139	86–134
12	76–154	79–151	84–146	89–141
13	79–161	82–158	88–152	92–148
14	81–169	85–165	91–159	96–154
15	84–176	88–172	94–166	99–161

$n_X = 11$

n_Y	.005	.01	.025	.05
11	87–166	91–162	96–157	100–153
12	90–174	94–170	99–165	104–160
13	93–182	97–178	103–172	108–167
14	96–190	100–186	106–180	112–174
15	99–198	103–194	110–187	116–181

$n_X = 12$

n_Y	.005	.01	.025	.05
12	105–195	109–191	115–185	120–180
13	109–203	113–199	119–193	125–187
14	112–212	116–208	123–201	129–195
15	115–221	120–216	127–209	133–203

$n_X = 13$

n_Y	.005	.01	.025	.05
13	125–226	130–221	136–215	142–209
14	129–235	134–230	141–223	147–217
15	133–244	138–239	145–232	152–225

$n_X = 14$

n_Y	.005	.01	.025	.05
14	147–259	152–254	160–246	166–240
15	151–269	156–246	164–256	171–249

$n_X = 15$

n_Y	.005	.01	.025	.05
15	171–294	176–289	184–281	192–273

Source: Adaptation of Table 1, L. R. Verdooren, "Extended Tables of Critical Values for Wilcoxon's Test Statistic," *Biometrika*, **50**, 177–186 (1963), with permission of the author and the Biometrika Trustees.

TABLE K *Critical Values for the Smaller of* R_+ *or* R_- *for the Wilcoxon Signed-Ranks Test*

	LEVELS OF SIGNIFICANCE FOR A ONE-TAILED TEST			
	.05	.025	.01	.005
	LEVELS OF SIGNIFICANCE FOR A TWO-TAILED TEST			
NUMBER OF PAIRS	.10	.05	.02	.01
4	—			
5	0	—		
6	2	0	—	
7	3	2	0	—
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	10	7
13	21	17	12	10
14	25	21	16	13
15	30	25	19	16

Source: Constance van Eeden, Nonparametric Introduction to Statistics, 1st edition, p. 221, © 1968. Reprinted by permission of Pearson Education, Inc., Upper Saddle River, NJ.

REFERENCES

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science,* 8, 12–15.

Adams, E. W. (1966). On the nature and purpose of measurement. *Synthese,* 16, 125–129.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

American Psychological Association. (2001). *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

Anderson, C. A., et al. (2003). The influence of media violence on youth. *Psychological Science in the Public Interest*, 4, 81–110.

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *Behavior Analyst,* 22, 93–97.

Bandura, A., Ross, D. A., & Ross, S. A. (1963). Imitation of film-mediated aggressive models. *Journal of Abnormal and Social Psychology,* 66, 3–11.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society,* 53, 370–418.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10, 389–396.

Berger, J. D., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist,* 76, 159–165.

Best, J. (2001). *Damned Lies and Statistics*. Berkeley, CA: University of California Press.

Best, J. B. (1979). Item difficulty and answer changing. *Teaching of Psychology,* 6, 228–230.

Bialik, C. (2008). Polls foresaw future, which looks tough for polling. *The Wall Street Journal*, Nov. 6, A16.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory *F* when variances are heterogeneous. *British*

Journal of Mathematical and Statistical Psychology, 40, 26–42.

BOLLES, R. C. (1988). Why you should avoid statistics. *Biological Psychiatry,* 23, 79–85.

BONEAU, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin,* 57, 49–64.

BOWERS, T. G., & CLUM, G. A. (1988). Relative contribution of specific and nonspecific treatment effects: Meta-analysis of placebo-controlled behavior therapy research. *Psychological Bulletin,* 103, 315–323.

BRADLEY, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice Hall.

BRADLEY, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology,* 31, 144–152.

BRADY, J. V., PORTER, R. W., CONRAD, D. G., & MASON, J. W. (1958). Avoidance behavior and the development of gastroduodenal ulcers. *Journal of the Experimental Analysis of Behavior,* 1, 69–72.

BROWN, K. S. (1995). Testing the most curious subject—oneself. *Scientist,* 9(24), 1.

BUSHMAN, B. J., & ANDERSON, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science*, 20, 273–277.

BYRNE, W. L., ET AL. (1966). Memory transfer. *Science,* 153, 658–659.

CAMILLI, G., & HOPKINS, K. D. (1978). Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. *Psychological Bulletin,* 85, 163–167.

CARNAGEY, N. L., ANDERSON, C. A., & BUSHMAN, B. J. (2007). The effect of video game violence on physiological desensitization to real-life violence. *Journal of Experimental Social Psychology*, 43, 489–496.

CHERNOFF, H., & SAVAGE, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics,* 29, 972–999.

CHOW, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences,* 21, 169–239.

CHRISTENSEN, P. N., & WOOD, W. (2006). Effects of media violence on viewers' aggression in unconstrained social interaction. In R. Preiss, B. M. Gayle, N. Burrell, M. Allen, & J. Bryant (Eds.), *Mass Media Research: Advances through Meta-Analysis* (pp. 145–168). Mahwah, NJ: Erlbaum.

CLEVELAND, W. S. (1995). *Visualizing Data*. Summit, NJ: Hobart Press.

CLEVELAND, W. S., & MCGILL, R. (1985). *The Elements of Graphing Data*. Belmont, CA: Wadsworth.

COHEN, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 65, 145–153.

COHEN, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp. 95–121). New York: McGraw-Hill.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist,* 45, 1304–1312.

COHEN, J. (1992). A power primer. *Psychological Bulletin,* 112, 155–159.

COHEN, J. (1994). The earth is round ($p < .05$). *American Psychologist,* 49, 997–1003.

COHEN, P. (1988). Are statistics necessary? *Biological Psychiatry,* 23, 1–2.

COLE, K. C. (1985). Is there such a thing as scientific objectivity? *Discover,* 6(9), 98–99.

CONOVER, W. J. (1974). Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables. *Journal of the American Statistical Association,* 69, 374–376.

CORMACK, R. S., & MANTEL, N. (1990). Doubt and certainty in statistics. *Journal of the Royal Society of Medicine,* 83, 136–137.

CRISLER, M. C., BROOKS, J. O., OGLE, J. H., GUIRL, C. D., ALLURI, P., & DIXON, K. K. (2008). The effect of wireless communication and entertainment devices on simulated driving performance. *Transportation Research Record: Journal of the Transportation Research Board*, 2069, 48–54.

CUMMING, G. (2008). Replication and *p* intervals. *Perspectives on Psychological Science*, 3, 286–300.

CUMMING, G., et al. (2007). Statistical reform in psychology. Is anything changing? *Psychological Science*, 18, 230–232.

CURRAN-EVERETT, D. (2000). Multiple comparisons: Philosophies and illustrations. *American Journal of Physiology,* 279, R1–R8.

DAVISON, M. L., & SHARMA, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin,* 104, 137–144.

DAVISON, M. L., & SHARMA, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin,* 107, 394–400.

DAWES, R. B. (1975). Graduate admission variables and future success. *Science,* 187, 721–723.

DEBUONO, B. A., ZINNER, S. H., DAAMEN, M., & MCCORMACK, W. M. (1990). Sexual behavior of college women in 1975, 1986, and 1989. *New England Journal of Medicine,* 322, 821–825.

DELUCHHI, K. L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin,* 94, 166–176.

DEWS, P. B. (1955). Studies on behavior. I. Differential sensitivity to pentobarbital of pecking performance in pigeons depending on the schedule of reward. *Journal of Pharmacology and Experimental Therapeutics,* 113, 393–401.

DIACONIS, P., & EFRON, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248, 116–130.

DIACONIS, P., & MOSTELLER, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association,* 84, 853–861.

DINGFELDER, S. F. (2004, December). The significance of null. *Monitor on Psychology*, 35, 24–25.

DOLL, R. (1955). Etiology of lung cancer. *Advances in Cancer Research,* 3, 1–50.

DREW, B., & WATERS, J. (1985). Video games: Utilization of a novel strategy to improve perceptual-motor skills in non-institutionalized elderly. *Proceedings and Abstracts of the Eastern Psychological Association,* 5, 56.

DUNN, O. J. (1964). Multiple comparisons using rank sums. *Technometrics,* 6, 241–252.

EBY, D., & KOSTYNIUK, L. (2003). *Driver Distraction and Crashes: An Assessment of Crash Databases and Review of the Literature*. Publication UMTRI-2003-12. University of Michigan Transportation Research Institute, Ann Arbor, 2003.

EDEN, D. (2003). Self-fulfilling prophecies in organizations. In J. Greenberg (Ed.), *Organizational Behavior* (2nd ed., pp. 91–122). Mahwah, NJ: Erlbaum.

EDEN, D., & ZUK, Y. (1995). Seasickness as a self-fulfilling prophecy: Raising self-efficacy to boost performance at sea. *Journal of Applied Psychology*, 80, 628–635.

EDGINGTON, E. S. (1987). *Randomization Tests* (2nd ed.). New York: Marcel Dekker.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.

ENGEL, R., & BENSON, R. C. (1968). Estimate of conceptual age by evoked response activity. *Biologia Neonatorum,* 12, 201–213.

EPSTEIN, J. L., & MCPARTLAND, J. M. (1976). The concept and measurement of the quality of school life. *American Educational Research Journal,* 13, 15–50.

FISCHL, M. A., ET AL. (1987). The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine,* 317, 185–191.

FISHER, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society,* 98, 39–54.

FISHER, R. A. (1949). *The Design of Experiments* (5th ed.). London: Oliver & Boyd.

FISHER, R. A. (1960). *The Design of Experiments* (7th ed.). Edinburgh, Scotland: Oliver & Boyd.

FLEISS, J. L. (1969). Estimating the magnitude of experimental effects. *Psychological Bulletin,* 72, 273–276.

FLEISS, J. L. (1981). *Statistical Methods for Rates and Proportions*. New York: Wiley.

FOLTZ, E. L., & MILLETT, F. E. (1964). Experimental psychosomatic disease states in monkeys. 1. Peptic "ulcer-executive" monkeys. *Journal of Surgical Research,* 4, 445–453.

FOX, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

FRANKLIN, R. D., ALLISON, D. B., & GORMAN, B. S. (Eds.) (1997). *Design and Analysis of Single-Case Research*. Mahwah, NJ: Erlbaum.

FREEDMAN, D., PISANI, R., & PURVES, R. (1978). *Statistics*. New York: Norton.

GALTON, F. (1889). *Natural Inheritance*. London: Macmillan.

GAMES, P. A., & KLARE, G. R. (1967). *Elementary Statistics. Data Analysis for the Behavioral Sciences.* New York: McGraw-Hill.

GARFIELD, E. (1991, October 28). Meta-analysis and the metamorphosis of the scientific literature review. *Current Contents,* 43, 5–8.

GEARY, R. C. (1947). Testing for normality. *Biometrika,* 34, 209–242.

GELMAN, A. (1998). Some class-participation demonstrations for decision theory and Bayesian statistics. *American Statistician,* 52, 167–174.

GIGERENZER, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kruger, G. Gigerenzer, & M. S. Morgan (Eds.), *The Probabilistic Revolution* (vol. 2, pp. 11–33). Cambridge, MA: Bradford/MIT Press.

GLASS, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Research,* 5, 3–8.

GLASS, G. V., & HOPKINS, K. D. (1996). *Statistical Methods in Education and Psychology* (3rd ed.). Boston, MA: Allyn and Bacon.

GLASS, G. V., PECKHAM, P. D., & SANDERS, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research,* 42, 237–288.

GREENHOUSE, S. W. (1990). Comment. *Statistics in Medicine,* 9, 371–372.

GUTTMACHER, A. F. (1973). *Pregnancy, Birth, and Family Planning*. New York: New American Library.

HAGEN, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist,* 52, 15–24.

HARRIS, R. J. (1997). Significance tests have their place. *Psychological Science,* 8, 8–11.

HARTWIG, F., & DEARING, B. E. (1979). Exploratory data analysis. In J. L. Sullivan (Ed.), *Quantitative Applications in the Social Sciences*. Beverly Hills, CA: Sage.

HAVILAND, M. G. (1990). Yates' correction for continuity and the analysis of 2×2 contingency tables. *Statistics in Medicine,* 9, 363–367.

HAYS, W. L. (1994). *Statistics* (5th ed.). New York: Holt, Rinehart, & Winston.

HEARNSHAW, L. S. (1979). *Cyril Burt, Psychologist.* Ithaca, NY: Cornell University Press.

HEDGES, L. V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics,* 6, 107–128.

HEDGES, L.V., & OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic.

HEDGES, L.V., & PIGOTT, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203–217.

HEDGES, L. V., & VEVEA, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics,* 21, 299–332.

HOAGLIN, D. C., MOSTELLER, F., & TUKEY, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

HOCK, R. R. (2006). *Forty Studies that Changed Psychology: Explorations into the History of Psychological Research* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

HOLLAND, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

HOPKINS, B. L., COLE, B. L., & MASON, T. L. (1998). A criticism of the usefulness of inferential statistics in applied behavior analysis. *Behavior Analyst,* 21, 125–137.

HOPKINS, K. D., GLASS, G. V., & HOPKINS, B. R. (1987). *Basic Strategies for the Behavioral Sciences* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

HOSMER, D. W., & LEMESHOW, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

HOWSON, C., & URBACH, P. (1991). Bayesian reasoning in science. *Nature,* 350, 371–374.

HUFF, D. (1993). *How to Lie with Statistics*. New York: Norton.

HUNTER, J. E. (1997). Needed: A ban on the significance test. *Psychological Science,* 8, 3–7.

JASIENSKI, M. (1996, March 4). Wishful thinking and the fallacy of single-subject experimentation. *Scientist,* 10.

JOHNSON, J. G., COHEN, P., SMAILES, E. M., KASEN, S., & BROOK, J. S. (2002). Television viewing and aggressive behavior during adolescence and adulthood. *Science*, 295, 2468–2471.

JUDGE, T. A., THORESEN, C. J., BONO, J. E., & PATTON, G. K. (2001). The job satisfaction-job performance relationships: A qualitative and quantitative review. *Psychological Bulletin*, 127, 376–407.

KAPLAN, R. M., & SACCUZZO, D. P. (1989). *Psychological Testing: Principles, Applications, and Issues* (2nd ed.). Pacific Grove, CA: Brooks/Cole.

KEHLE, T. J., BRAY, M. A., CHAFOULEAS, S. M., & KAWANO, T. (2007). Lack of statistical significance. *Psychology in the Schools*, 44, 417–422.

KEPPEL, G. (1991). *Design and Analysis: A Researcher's Handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

KEREN, G., & LEWIS, C. (1979). Partial omega squared for ANOVA designs. *Educational and Psychological Measurement,* 39, 119–128.

KIRK, R. E. (1990). *Statistics: An Introduction*. Fort Worth, TX: Holt.

KIRK, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

KIRK, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement,* 56, 746–759.

KOLATA, G. (1988). Theorist applies computer power to uncertainty in statistics. *New York Times*, Nov. 11, C1, C6.

KRUEGER, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist,* 56, 16–26.

LANG, A. R., GOECKNER, D. J., ADESSO, V. I., & MARLATT, G. A. (1975). Effects of alcohol and aggression in male social drinkers. *Journal of Abnormal Psychology,* 84, 508–518.

LEVINE, T. R. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, 76, 286–302.

LOFTUS, G. R. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers,* 25, 250–256.

LOFTUS, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science,* 5, 161–171.

MACKINTOSH, N. J. (Ed.) (1995). *Cyril Burt: Fraud or Framed?* Oxford: Oxford University Press.

MAHONEY, M. J. (1976). *Scientist as Subject. The Psychological Imperative*. Cambridge, MA: Ballinger.

MANN, C. (1990). Meta-analysis in the breech. *Science,* 249, 476–480.

MANN, H. B., & WHITNEY, D. R. (1947). On a test of whether one or two random variables is stochastically larger than the other. *Annals of Mathematical Statistics,* 18, 50–60.

MANSON, J. E., et al. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine*, 349, 523–534.

MANTEL, N. (1990). Comment. *Statistics in Medicine,* 9, 369–370.

MANTEL, N., & GREENHOUSE, S. W. (1968). What is the continuity correction? *American Statistician,* 22, 27–30.

MATARAZZO, J. D. (1972). *Wechsler's Measurement and Appraisal of Adult Intelligence: Fifth and Enlarged Edition*. New York: Oxford University Press.

MAXWELL, S. E., & DELANEY, H. D. (1985). Measurement scales and statistics: An examination of variable validity. *Psychological Bulletin,* 97, 85–93.

MAY, R. B., MASSON, M. E. J., & HUNTER, M. A. (1997). *Application of Statistics in Behavioral Research* (2nd ed.). New York: Harper Row.

MCCALL, R. B. (1977). Childhood IQ's as predictors of adult educational and occupational status. *Science,* 197, 482–483.

MCCANN, I. L., & HOLMES, D. S. (1984). Influence of aerobic exercise on depression. *Journal of Personality and Social Psychology,* 46, 1142–1147.

MCHUGH, R. B. (1963). Comment on "Scales and statistics: Parametric and non-parametric." *Psychological Bulletin,* 60, 350–355.

MICCERI, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin,* 105, 156–166.

MICHAEL, W. B. (1966). An interpretation of the coefficients of predictive validity and of determination in terms of the proportions of correct inclusions or exclusions in cells of a fourfold table. *Educational and Psychological Measurement,* 26, 419–424.

MILLIGAN, G. W., WONG, D. S., & THOMPSON, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin,* 101, 464–470.

MOHER, D., DULBERG, C. S., & WELLS, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association,* 272, 122–124.

MOHER, D., & PHAM, B. (1999). Meta-analysis: An adolescent in need of evidence and a watchful eye. *Annals of Medicine,* 31, 153–155.

MORGAN, D. L., & MORGAN, R. K. (2001). Single-participant research design. Bringing science to managed care. *American Psychologist,* 56, 119–127.

MUTHÉN, B. O. (1987). Response to Freedman's critique of path analysis: Improve credibility by better methodological training. *Journal of Educational Statistics,*12, 178–184.

NATELSON, B. (1976). The "executive" monkey revisited. Paper presented at the Symposium on Nerves and the Gut, Philadelphia.

NEAVE, H. R., & GRANGER, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics,* 10, 509–522.

NELDER, J. A. (1988). Response to Berger and Berry. *American Scientist,* 76, 431–432.

NEYMAN, J., & PEARSON, E. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika,* 20A, 175–240.

NEYMAN, J., & PEARSON, E. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika,* 20A, 263–294.

NUNNALLY, J. C. (1978). *Psychometric Theory*. New York: McGraw Hill.

PERONE, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *Behavior Analyst,* 22, 109–116.

PETERSON, C. (2009). Minimally sufficient research. *Perspectives on Psychological Science*, 4, 7–9.

PILIAVIN, I. M., RODIN, J., & PILIAVIN, J. A. (1969). Good Samaritan: An underground phenomenon. *Journal of Personality and Social Psychology,* 13, 289–299.

POLANYI, M. (1961). The unaccountable element in science. *Transactions of the Bose Research Institute,* 24, 175–184.

PRUZEK, R. M. (1997). An introduction to Bayesian inference and its applications. In L. Harlow, S. A. Mulaik, & J. Steigen (Eds.), *What If There Were No Significance Tests?* (pp. 285–316). Hillsdale, NJ: Erlbaum.

ROGAN, J. D., & KESELMAN, H. J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investment via a coefficient of variation. *American Educational Research Journal,* 14, 493–498.

ROSCOE, J. T., & BYERS, J. A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association,* 66, 755–759.

ROSENTHAL, R., & DIMATTEO, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology,* 52, 59–82.

ROSNOW, R. L., & ROSENTHAL, R. (1989a). Definition and interpretation of interaction effects. *Psychological Bulletin,* 105, 143–146.

ROSNOW, R. L., & ROSENTHAL, R. (1989b). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist,* 44, 1276–1284.

ROSNOW, R. L., & ROSENTHAL, R. (1991). If you're looking at cell means, you're looking at only the interaction (unless all main effects are zero). *Psychological Bulletin,* 110, 574–576.

ROSNOW, R. L., & ROSENTHAL, R. (1995). "Some things you learn aren't so": Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science,* 6, 3–9.

ROSNOW, R. L., ROSENTHAL, R., & RUBIN, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science,* 11, 446–453.

Rosopa, P. J., & Stone-Romero, E. F. (2008). Problems with detecting assumed mediation using the hierarchical multiple regression strategy. *Human Resource Management Review*, 18, 294–310.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology,* 58, 646–656.

Rotton, J., & Kelly, I. (1985). Much ado about the full moon: A meta-analysis of lunar-lunacy research. *Psychological Bulletin,* 97, 286–306.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese,* 16, 170–233.

Samelson, F. (1996). He didn't? Yes, he did (probably)! *Contemporary Psychology,* 41, 1177–1179.

San José, B., van de Mheen, H., van Oers, J. A. M., Mackenbach, J. P., & Garretsen, H. F. L. (1999). The U-shaped curve: Various health measures and alcohol drinking patterns. *Journal of the Study of Alcohol,* 60, 725–731.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin,* 111, 352–360.

Scarr, S. (1985). Constructing psychology: Making facts and fables for our times. *American Psychologist,* 40, 499–512.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist,* 47, 1173–1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods,* 1, 115–129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin,* 105, 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology,* 64, 1290–1305.

Shapiro, S. (1994). Meta-analysis/shmeta-analysis. *American Journal of Epidemiology,* 140, 771–778.

Sidman, M. (1960). *Tactics of Scientific Research*. New York: Basic Books.

Siegel, S. S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd ed.). New York: McGraw Hill.

Sinclair, J. D. (1988). Multiple *t*-tests are appropriate in science. *Trends in Pharmacological Science,* 9, 12–13.

Skinner, B. F. (1966). What is the experimental analysis of behavior? *Journal of the Experimental Analysis of Behavior,* 9, 213–218.

Slack, W. V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review,* 50, 154–175.

Smith, L. D., Best, L. A. Stubbs, D. A., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge. The role of graphs and tables in hard and soft psychology. *American Psychologist*, 57, 749–761.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist,* 32, 752–760.

Steering Committee Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine,* 318, 262–264.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association,* 54, 30–34.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science,* 103, 677–680.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin,* 105, 147–155.

Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods*, 11, 326–352.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson.

Taylor, S. E. (1991). *Health Psychology.* New York: Random House.

Terwilliger, J. S. (1978). Stevens' scales of measurement: A critique and an alternative. Unpublished manuscript, Department of Educational Psychology, University of Minnesota, Minneapolis.

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin,* 99, 90–99.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin,* 96, 394–401.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin,* 76, 105–110.

URY, H. (1967). In response to Noether's letter, "Needed—a new name." *American Statistician,* 21, 53.

VAUGHAN, G. M., & CORBALLIS, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin,* 72, 204–223.

WACHTER, K. W. (1988). Disturbed by meta-analysis? *Science,* 241, 1407–1408.

WAINER, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods,* 4, 212–213.

WEISS, D. J. (1989). An experiment in publication: Advance publication review. *Applied Psychological Measurement,* 13, 1–7.

WEISS, J. M. (1971). Effects of coping behavior with and without a feedback signal on stress pathology in rats. *Journal of Comparative and Physiological Psychology,* 77, 22–30.

WHITLEY, B. E., JR. (1992). Units of analysis, measurement scales, and statistics: A comment on Kerwin and Shaffer. *Personality and Social Psychology Bulletin,* 18, 680–684.

WILCOX, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53, 300–314.

WILCOX, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (2nd ed.). Burlington, MA: Elsevier.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,* 1, 80–83.

WILCOXON, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics,* 3, 119–122.

WILKINSON, L. (1988). *SYSTAT: The System for Statistics*. Evanston, IL: SYSTAT.

WILKINSON, L., AND THE TASK FORCE ON STATISTICAL INFERENCE. (1999). Statistical methods in psychology journals. Guidelines and explanations. *American Psychologist,* 54, 594–604.

WILLIAMS, M., LECLUYSE, K., & ROCK-FAUCHEUX, A. (1992). Effective interventions for reading disabilities. *Journal of the American Optometric Association,* 63, 411–417.

WINER, B. J., BROWN, D. R., & MICHELS, K. M. (1991). *Statistical Principles in Experimental Design* (3rd ed.). New York: McGraw Hill.

YATES, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (supplement), 1, 217–235.

This page intentionally left blank

INDEX

A
Abscissa, 34
Addition theorem of probability, 158
Alpha (α), 186, 219–220
Alternative hypothesis, 185–186
  in ANOVA, 300
  in assumption-free tests, 382
  in chi-square, 364, 371
  directional, 190, 223, 242
  nondirectional, 185–186, 223
  in *t* tests of difference between two means, 234, 262
  in *z* and *t* tests of single means, 185–186, 223
Analysis of variance (ANOVA), 300
  vs. *a priori* comparisons, 319–320
  comparisons with *t*, 310–311
  logic of, 300–306
  one-way, 299–333, 335–340
    alternative hypothesis, 300
    assumptions about, 315–316, 328, 352
    between-groups variation, 302–303, 306
    degrees of freedom in, 305, 325
    effect size and, 316–317, 328
    F distribution and, 306
    F ratio and, 305–307
    grand mean and, 302
    inherent variations and, 303, 305
    null hypothesis and, 300–301
    post hoc comparisons and, 318–319
    power and, 317
    raw score formulas for, 311–313, 325
    for repeated measures, 323–328
    summary table for, 307, 326
    sum of squares and, 302–304, 324
    treatment conditions for, 300
    treatment effect and, 301
    within-groups variation and, 301, 305
  vs. planned comparisons, 320–323
  for repeated measures, 323–328
  two-way, 335–358
    assumptions of, 352
    cells, 336
    columns variation, 343–344
    degrees of freedom, 346, 356
    effect size and, 349–350
    F tests, 346–347, 356
    interaction and, 338–341
    interpreting outcomes, 348–349
    main effects of, 336–338
    mixed within-subjects design, 352–357
    planned comparisons and, 352
    rows variation and, 343
    rows X columns (interaction) and, 338–340
    simple effects and, 341
    sum of squares and, 341–345
    total sum of squares and, 341–342
    variance estimates and *F* tests, 346–347
    within-cell variation and, 342–343
"And" rule of probability, 159
ANOVA, *see* Analysis of variance (ANOVA); ANOVA tables
ANOVA tables
  one-way, 307, 326
  two-way, 347, 356
Apparent limits of a score interval, 22
Applied statistics, 4–6
A priori comparisons, 320, 352
Assumption-freer methods, 381–401
  bootstrap method, 385–386
  Friedman's rank test, 397–399
  Kruskal-Wallis test, 395–397
  Mann-Whitney *U* test, 386–390
  null hypothesis in, 382
  randomization tests and, 382–385
  rank-order tests and, 385
  sign test and, 390–392
  Spearman's rank-order correlation coefficient and, 112–113, 279
  Wilcoxon rank-sum test and, 387
  Wilcoxon signed-ranks test and, 392–394
Asymmetrical distribution, 43, 56–57
Asymptotic curve, 85

B
Bar diagram, 38–40
Bayesian statistics, 289–90
Bell-shaped distribution, 44
Beta (β), 220
Between-groups variance estimate, 294, 305–306
Bimodal distribution, 44
Binomial distribution, 159–162, 440–442
Binomial expansion, 161–162
Bivariate distributions, 99–102, 370
Bootstrap method, 385–386

C

- Categorical data, 18, 364
- Cell, 336, 370
- Cell frequency, 371
- Central limit theorem, 176, 251, 369
- Central tendency, 48
- Chi-square distribution, 366–370
- Chi-square statistic, 365
- Chi-square tests, 363–378
  - assumptions in use of, 370
  - contingency tables and, 372–374
  - degrees of freedom and, 366, 374
  - effect size and, 369, 374
  - expected frequencies of, 365–366, 372
  - Friedman's rank test and, 397
  - goodness-of-fit problems and, 364–370
  - null hypothesis in, 364, 368, 371
  - observed frequencies and, 365–366
  - power and, 377
  - sampling distributions and, 367
  - sign test and, 391
  - as test for independence, 370–372
  - Yates' correction and, 375
- Class intervals, 19
- Coefficient of determination, 151, 279
- Coefficient of nondetermination, 148
- Cohen's d, 217, 246, 267, 289, 292
- Concordance coefficient, 399
- Conditional probability, 165
- Confidence coefficient, C, 285
- Confidence intervals, 281–297
  - advantages of, 286–287
  - evaluating results of, 288–290
  - versus hypothesis testing, 286
  - of ρ, 295–296
  - of μ, 283–286
  - of $\mu_x - \mu_y$, 290–295
- Confounding variables, 5
- Constant, 6
- Contingency tables, 371
- Continuous variable, 6
- Correlation
  - and cause-and-effect relationship, 114–115
  - meaning of degree of, 103–104
  - negative, 103
  - none, 103
  - phi coefficient (ϕ), 295–296
  - positive, 102
  - Spearman's rank-order, 112–113
- Correlation coefficient (r and r_s), 106–107. *See also* Regression equation
  - cautions concerning, 116–119
  - Fisher's z' transformation, 277–278
  - formulas for
    - using deviation scores, 107
    - using IBM SPSS, 109–111
    - using raw scores, 107–109
  - hugging principle, 144
  - inference about, 274–280
  - product-moment, 106
  - range of talent, sensitivity to, 116
  - score transformation (linear) effects of, 115–116
  - Spearman's rank-order, 112–113
  - values required for different levels of significance, 452–453
- Cramer's phi, 375–376
- Critical value(s), 188
- Cross-products, 106
- Cumulative frequency distributions, 24–25, 41–42
- Cumulative percentage curve, 41–42
- Cumulative percentage distribution, 25–26, 41–42
- Curvilinear relationship, 101–102, 115

D

- d, 217, 246, 267, 289, 292
- Degrees of freedom,
  - defined, 197
  - for chi-square tests, 366, 374
  - in one-way ANOVA, 305, 325
    - t test, 198, 239, 261, 276
  - in two-way ANOVA, 346, 356
- Dependent-samples design, 259
  - problems with using, 268–270
- Dependent variable, 5
- Descriptive statistics, 3
- Determination, coefficient of, 151, 279
- Deviational measures, 64–66
- Deviation score, 52, 64
- Dichotomous decision making, 282
- Dichotomous observations, 160
- Difference scores, 261
- Directional (one-tailed) test, 190, 233, 242
- Discontinuous distributions, 117
- Discrete variable, 6
- Disordinal interaction, 340
- Distribution
  - binomial, 159–162
  - bivariate, 99–102
  - of chi-square, 366–370
  - cumulative frequency, 24–25
  - discontinuous, 117
  - frequency, 17–30
  - normal, 85, 92–93
  - open-ended, 55
  - probability, 159–162
  - random sampling distribution of the mean, 172–177, 185
  - relative frequency, 22–24
  - sampling, 169
  - shape of, 43–45
  - skewed, 44
  - student's distribution of t, 196
  - t distribution, 195–196
- Distribution-free methods, 382

E

- Effect size
  - d family, 217, 246, 267, 289, 292
  - r family, 248, 316, 349–350
  - definition, 216–217
  - for chi-square, 369, 374
  - for correlation, 279
  - for Friedman's rank test, 399
  - for Kruskal-Wallis test, 397
  - for Mann-Whitney U test, 389
  - for one-way ANOVA, 316–317, 328
  - for t, single means, 217
  - for t, two dependent means, 267–268
  - for t, two independent means, 246–249
  - for two-way ANOVA, 349–350
  - for Wilcoxon signed-ranks test, 394
  - versus p-values, 216–218
- Empirical probability, 157
- Epsilon-squared, 397
- Equally-likely model of probability, 157
- Error
  - in estimating Y from X, 136–138
- Estimated standard error
  - for dependent samples, 260–261
  - of the difference between two means, 237–238
  - of estimate, 132–134
  - of the mean, 193–195
- Estimation, 282–283. *See also* Confidence intervals
- Eta-squared, 316, 328, 349
- Event, 156
  - independent, 159, 166
- Expected frequency, 364, 372
- Expected value, 175
- Experimental designs
  - factorial, 336
  - matched-pairs investigations, 259
  - matched-subjects design, 259
  - repeated-measures, 259, 324, 352
  - single-subject research design, 192–193

F

- f (frequency), 21
- Factorial design, 336
- Factors, 300, 336
- F distribution, 306
- Fisher's z', 277–278
- F ratio
  - for one-way ANOVA and, 306, 325
  - for repeated measures ANOVA, 325
  - two-way ANOVA and, 346–347, 356
- Frequency
  - expected, 364–366
  - observed, 364
- Frequency distribution, 17–30
  - bell-shaped, 44, 82–95
  - bimodal, 44
  - bivariate, 370
  - cumulative, 24–25
  - graphical representations of, 33–45
  - grouped, 21–22
  - J-shaped distribution, 44
  - normal, 82–95
  - open-ended distribution, 55
  - rectangular, 44
  - relative, 22–24

  - shape of, 43–45
  - skewed, 44
- Frequency polygon, 36–37
- Friedman's rank test, 397–399

G

- *g*, 217, 246, 267
- Gambler's fallacy, 159
- Generalizing results, 229, 405
- Generating event, 156
- Glass rank biserial correlation coefficient, 389
- Goodness-of-fit problems, 364. *See also* Chi-square tests
- Grand mean, 302
- Graphs, 33–45
  - bar diagram, 38–40, 43
  - cumulative percentage curve, 41–42
  - factors affecting shape of, 42–45
  - frequency polygons, 36–38
  - histograms, 34–38
  - pie charts, 40
  - scatter diagrams, 99–102
  - zero point on, 34
  - Greek letter symbols, 416–417
- Grouped frequency distribution, 21–22
- Grouped scores, 19–20
- Grouping error, 19

H

- Harmonic mean, 319
- Hedges' *g*, 217, 246, 267
- Histograms, 34–36
- Homogeneity of variance, 237, 254, 304
- Homoscedasticity, 138
- Hugging principle, 144
- Hypothesis
  - alternative, 185–186, 233
    - in ANOVA, 300
    - directional, 190, 223, 242
    - nondirectional, 190
  - defined, 185
  - null, 185–186
    - in ANOVA, 300, 337
    - in assumption-freer tests, 382
    - in chi-square tests, 364, 371
    - in correlation, 276
    - defined, 185–186
  - of independence, 371
  - main effect, 337
  - one-way analysis of variance and, 301–302
  - rejection of, 189
  - in *t* and *z* tests of single means, 185–186
  - in *t* tests of difference between two means, 233, 261
- Hypothesis testing
  - with assumption-freer tests, 382
  - with categorical data (chi-square test), 364
  - vs. confidence intervals, 283–286
  - of correlation coefficients, 274–279
  - logic of, 184–185
  - of no difference between two independent means, 233–253
  - one-tailed tests, 190–191, 242
  - power of the test and, 220–221, 249
  - region of rejection, 187, 368
  - region of retention, 187
  - single mean, 184–209
  - statistical vs. practical differences, 214–215
  - with three or more groups (ANOVA), 299–332
  - two dependent means, 259–270
  - two independent means, 233–254
  - two-tailed tests, 190–191
  - Type I and Type II errors, 218–220

I

- *i* (interval width), 21
- IBM SPSS, 13–14
  - for ANOVA, 313–315, 328–330, 350–352, 357–358
  - for correlation, 109–111
  - for regression, 134–136
  - for standard deviation, 68–71
  - for *t*, dependent groups, 265–267
  - for *t*, independent groups, 242–245
  - for *t*, single mean, 204–206
- Independence
  - null hypothesis for, 364, 371
  - in a probability sense, 159, 166
  - between two variables, 370–372
- Independent events, 159, 166
- Independent samples, 233
- Independent variable, 5
- Inferential statistics, 3–4, 169, 184
- Inherent variation, 303, 305
- Interaction, 338–340
  - disordinal, 340
  - importance of, 340–341
  - ordinal, 340
- Interval(s), 19–20
  - estimates, 283
  - midpoint, 36
  - scale, 8

J

- J-shaped distribution, 44

K

- *k*
  - for number of categories, 364
  - for treatment conditions, 300
- Kruskal-Wallis test, 395–397
- Kurtosis, 44

L

- Least-squares criterion, 127
- Least-squares regression, vs. resistant line, 128
- Leptokurtic, 44
- Level of significance, 186, 206–207, 222–223
- choosing, 218–220
  - defined, 186
  - and power of tests, 222–223
  - versus *p*-values, 206–207
- Levels, of factors, 300, 336
- Linear interpolation, 27
- Linearity, assumption of, 127
- Linear transformation, 57, 74, 115
- Line of best fit, 127

M

- Main effect, 336–338
- Mann-Whitney *U* test, 386–390
- Marginal totals, 370
- Matched-pairs investigations, 259
- Matched-pairs rank biserial correlation coefficient, 394
- Matched-subjects design, 259, 268
- Mathematics
  - review of basic skills, 407–414
  - test of basic skills, 414–415
- *Mdn* (symbol for median), 49–50
- Mean, 50–51
  - as balance point, 52, 58
  - of distribution of scores, converted to *z* scores, 76
  - estimated standard error of the difference between, 237
  - grand mean, 302
  - hypothesis testing about, 184
  - of population, 51
  - properties of, 52–54
  - random sampling distribution of, 172–178
  - regression line as, 128
  - regression toward, 147
  - of sample, 51
  - sampling distribution of difference between, 234–236
  - standard error of, 188–193
- Mean square between, 306
- Mean square error, 306
- Mean square within, 306
- Measurement
  - defined, 7
  - scales of, 7–9
- Measures of central tendency, 49–58
  - mean, 50–54
  - median, 49–50, 55
  - mode, 49, 51–52
  - in normal vs. skewed distribution, 56–57
- Measures of variability, 62–79
  - range, 21, 63
  - standard deviation, 66–74
  - variance, 64–65
- Median, 49–50
  - properties of, 55
- Mesokurtic, 44
- Meta-analysis, 225–226
- Minium's laws
  - First, 13, 340
  - Second, 66
  - Third, 171
- Mixed two-factor within-subjects ANOVA, 352–357

- *Mo* (symbol for mode), 49
- Mode, 49
  - properties of, 51–52
- Moment, 106
- Multiplication theorem of probability, 159
- Mutually exclusive outcomes, 158–159

N

- *N* (total number of cases in a population), 21
- *n* (total number of cases in a sample), 21
- Negative correlation, 103
- Negatively skewed distribution, 44, 56
- No correlation, 103
- Nominal scale, 7, 9
- Nondetermination, coefficient of, 148
- Nondirectional alternative, 190
- Nondirectional (two-tailed) test, 190, 233
- Nonparametric statistics, 382
- Nonsignificant results, 215
- Normal curve, 44, 82–95
  - areas under, corresponding to given values of z, 436–439
  - area under that falls between two known scores, 89–90
  - characteristics of, 85
  - equation of, 94
  - historical aspects of, 83–85
  - as model for real variables, 92
  - as model for sampling distributions, 92–93
  - standard, 87–92
  - standard scores and, 86–87
- Null hypothesis, 185–186
  - in ANOVA, 300–301
  - in assumption-freer tests, 382
  - in chi-square test, 364, 371
  - criteria for rejection/retention, 187–190
  - defined, 185
  - of independence, 371
  - one- and two-tailed tests and, 190, 223–224
  - rejection of, 186, 189
  - retention of, 186, 189
  - in tests of single means, 184–185

O

- Observed frequencies, 364
- Observer bias, 193
- Odds, 163, 376
- Odds ratio, 376
- Ogive curve, 41
- Omega, 369
- Omega-squared, 316–317
- Omnibus hypothesis, 300
- One-tailed test, 190–191, 223–224
  - power of the test and, 224
- One-way analysis of variance, 300–332
- Open-ended distribution, 55
- Order effect, 269
- Ordinal interaction, 340
- Ordinal scale, 7–8
- Ordinate, 34
- "Or" rule of probability, 158
- Orthogonal comparisons, 321–323
- Outlier, 55, 72, 385

P

- *P*
  - percentile, 25
  - proportion, 364
- Parameter, 4, 169
- Partial omega-squared, 349
- Partition of the sums of squares, 302–305, 324–327, 341–345, 353–356
- Partition of variance, 151
- Pearson's coefficient of correlation, 98–119, 143–153, 274–278
- Percentage curve, cumulative, 41–42
- Percentile, 25–26
- Percentile rank, 25–26, 77
- Phi coefficient (ø), 374
- Pie chart, 40
- Placebo effect, 193
- Planned *(a priori)* comparisons, 320–323, 352
- Platykurtic, 44, 197
- Point estimates, 283
- Polygon, frequency, 36–37
  - vs. histogram, 37–38
- Pooled estimate of population variance, 238
- Population, 4, 169
  - formula for mean of, 51
  - parameter, value of, 169
  - standard deviation of, 66
  - variance, pooled estimate of, 238
- Positive correlation, 102–103
- Positively skewed distribution, 44, 56
- *Post hoc* comparisons, 318–319, 347–348
- Power of a test, 220
  - ANOVA and, 317
  - and chi-square test of independence, 377
  - defined, 220
  - difference between two dependent means, 268
  - difference between two independent means, 249–250
  - factors affecting, 221–224
  - tests of a single mean, 220–229
- Power curves, 227
- Prediction, 98, 124–142. *See also* Regression
- Predictive error, 138–139
- Probability, 155–167
  - addition theorem of, 158
  - "and" rule of, 159
  - conditional, 165–166
  - defined, 156–157
  - empirical, 157
  - equally-likely model of, 157
  - of event, 156
  - independent events and, 159, 166
  - multiplication theorem of, 159
  - mutually exclusive outcomes and, 158
  - "or" rule of, 158
- Probability distribution, 160
  - binomial and, 159–162
  - defined, 160
- Probability samples, 170
- Product-moment correlation coefficient, 106
- *p*-value, 206–207, 216, 282

Q

- *Q* (semi-interquartile range), 63
- Qualitative data, organizing, 18–19, 38–40
- Qualitative variables, 7, 18, 38–45, 364
- Quantitative variables, 7–9
- Quartile points, 63

R

- *r*, 103, 106, 248. *See also* Correlation coefficient (r and r_s)
  - calculation of, 106–111
  - factors influencing, 144
  - interpretation of, 144–152
    - proportion of correct placements, 152
    - proportion of variance in Y associated with variation in X, 149–152
    - proportion of variation in Y not associated with variation in X, 148–149
  - as measure of effect size, 248
  - random sampling distribution of, 275–276
  - regression equation, 144–146
  - values of, 104
- r^2, 151, 279
- Random assignment, 251–253, 383
- Randomization tests, 382–385
- Random numbers, 171–172, 443–444
- Random sampling, 4, 168–182
  - as experimental control, 252–253
  - generalizing results and, 287–288
  - problems in selecting and in drawing conclusions, 229
  - vs. random assignment, 251–253
  - with replacement, 171
  - without replacement, 171, 180–181
- Random sampling distribution, 172–180
  - of chi-square, 367
  - of F, 306–307
  - of mean, difference between two sample means, 234–236
  - normal curve as model for, 92–93
  - of r, 275
  - of z', 277–278
- Range, 21, 63
  - properties of, 72
  - semi-interquartile range, 63, 72
- Range of talent, 116
- Rank-order tests, 385
- Ratio scale, 8–9
- Raw scores, 19
  - transforming to z scores, 75–76
- Real limits of a score, 22
- Rectangular distribution, 44

Region of rejection, 187
Region of retention, 187
Regression, 124–140, 143–152
  coefficient, 146
  with IBM SPSS, 134–136
  toward the mean, 147
  standard-score form of equation, 129
  of *Y* on *X,* 130–131
Regression equation, 126
  raw-score form of, 130–131
  standard-score form of, 129
Regression line, 126
  best-fit (least-squares), 127
  vs. resistant line, 128
  as running mean, 128
Rejecting the null hypothesis, 186, 189, 214–215
Rejection region, 187
Relative cumulative frequency distribution, 24–25, 41
Relative frequency distribution, 22–24, 160
Repeated-measures ANOVA, 323–330, 352–358
Repeated-measures design, 259, 324, 352
Research conclusion, 5, 200, 229, 405
Research question, 4, 403
Resistant line, 128
Retaining the null hypothesis, 186, 189, 214–215
Retention region, 187
Rho, Spearman's, 112

S
S (standard deviation of a sample), 66
Sample
  defined, 4, 169
  dependent, 259
  formula for mean of, 51
  independent, 233
  probability, 170
  standard deviation of, 66
Sample variance, calculation of, 64–71
Sampling. *See also* Random sampling
  random, 4, 170–171
  with replacement, 171
  without replacement, 171, 180
Sampling distribution. *See* Random sampling distribution
Scales of measurement, 7–9
  interval scale, 8
  nominal scale, 7
  ordinal scale, 7–8
  problems of statistical treatment and, 9–10
  ratio scale, 8–9
Scatter diagrams, 99–104
Scientific fraud, 11
Scores. *See* specific types
Score transformations, 57, 74, 115–116
Semi-interquartile range, 63, 72
Shape
  of distribution, 42–44
  of the sampling distribution, 175–176
Sigma, 51
Significant result, 186, 206–207
  power of the test and, 222
Sign test, 390–392
Simple effects, 341
Single-subject research design, 192–193
Skewed distribution, 44, 56
Slope of a line, 144
Spearman's rank-order correlation, coefficient (r_s), 112–113, 279
Spearman's rho, 112
Sphericity, assumption of, 328
SPSS, *see* IBM SPSS
SS (sum of squares), 65, 303–305, 324–325, 341–344, 353–357
Standard deviation, 67, 71
  calculation with IBM SPSS, 68–71
  calculation with raw-score method, 66–68
  of distribution expressed in *z* scores, 74–77
  of population, 66
  properties of, 72
  of sample, 66
  size of, 73
Standard error, 84, 175
  of a comparison, 320
  for dependent samples, 260–261
  of estimate, 131–133
  of the mean, 175, 193–195
  between two independent means, 236–238
Standardized normal curve, 82–95
  finding areas when score is known, 87–90
  finding scores when the area is known, 90–92
Standard scores, 74–77
  normal curve and, 82–95
  z-score formula for a value of $\bar{X}$, 178
Statistic(s)
  applied, 4–6
  defined, 3–4
  descriptive, 3
  inferential, 4
  lying with, 10–11
  unbiased, 71
Statistical conclusion, 5, 200, 229, 405
Statistical inference, 169, 184. *See also* Hypothesis testing; Inference
Statistically significant result, 186, 206–207, 214
Statistical question, 5, 403
Statistical tables, 435–458
Studentized range statistic, 318, 451
Student's distribution of *t,* characteristics of, 197–198
Summary table
  for one-way ANOVA, 307, 326
  for two-way ANOVA, 347, 356
Summation rules, 413–414
Sum of squares, 72, 302–305, 324–325, 353–356
Symbols
  English letter, 417–419
  Greek letter, 416–417
  and meanings of, 407–408
Symmetrical distribution, 56–57, 85
S_{YX}, 132–134
  error of prediction and, 144

T
Table of random numbers, 171–172, 443–444
t distribution, 196–198, 445–446
Theoretical relative-frequency distribution, 160
Treatment conditions, in one-way analysis of variance, 300–301
Treatment effect, 301
t test,
  comparison with *F,* 310
  formula for, 196, 202, 239, 260, 261
Tukey's *HSD* test, 318–319, 327, 348
Two-tailed test, 190–191, 223–224
Two-way analysis of variance (ANOVA)
  independent groups, 336–352
  mixed design, 352–358
Type I error, 218–220, 300
Type II error, 218–220

U
Unbiased estimator, 194, 208–209
Unbiased statistic, 71
Unimodel distribution, 85
U test. *See* Mann-Whitney *U* test

V
Variability, 61–79
  measures of, 62
  power of the test and, 222
  range, 63, 72
  semi-interquartile range, 63, 72
  standard deviation, 66–73
  variance, 64–66
Variables, 6–7
  categorical, 364
  chi-square as test for independence between, 370–377
  confounding, 5
  continuous, 6
  dependent, 5
  discrete, 6
  independent, 5
Variance, 64–66
  assumption of homogeneity of, 237, 304
  calculation with raw-score method, 66–68
  deviation-score method of calculating, 64–66
  in two-way analysis of, 335–360
  one-way analysis of, 300–332
  pooled estimate of population, 238

Variance, (*continued*)
  population, 65
  sample, 65
  in Y associated with variation in X, 149–152
Variation
  inherent, 303, 305
  within-groups, 301, 305, 342

W

Wilcoxon rank-sum test, 387
Wilcoxon signed-ranks test, 392–394
Within-cells sum of squares, 342
Within-groups variation, 301, 305, 342

X

X axis, 34

Y

Y', 127
Yates' correction for continuity, 375
Y axis, 34
Y intercept, 144

Z

Zero point on a graph, 34
z scores, 74–77. *See also* Standard scores
  areas under the normal curve corresponding to given values of z, 436–439
  formula for value of $\bar{X}$, 178
  z' transformations, Fisher's, 277–278

This page intentionally left blank

USEFUL FORMULAS

Here are some of the computing formulas that appear in the book. The reference is to the section number where each formula is introduced and defined.

Sample Mean	$\overline{X} = \frac{\sum X}{n}$	Sec. 4.3
Sum of Squares	$SS_X = \sum (X - \overline{X})^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$	Sec. 5.3
Sample Standard Deviation	$S_X = \sqrt{\frac{SS_X}{n}}$	Sec. 5.4
z Score from a Raw Score	$z = \frac{X - \overline{X}}{S_X}$ or $\frac{X - \mu_X}{\sigma_X}$	Sec. 5.10
Sum of Products of Deviation Scores (Raw-Score Formula)	$\sum (X - \overline{X})(Y - \overline{Y}) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}$	Sec. 7.7
Pearson Correlation Coefficient	$r = \frac{\sum (z_X z_Y)}{n} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{(SS_X)(SS_Y)}}$	Sec. 7.6–7.7
Spearman Rank-Order Correlation Coefficient	$r_S = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$	Sec. 7.9
Predicted z_Y from the Straight Line of Best Fit	$z'_Y = rz_X$	Sec. 8.3
Predicted *Y* from the Straight Line of Best Fit	$Y' = \left(r\frac{S_Y}{S_X}\right)X - \left(r\frac{S_Y}{S_X}\right)\overline{X} + \overline{Y}$	Sec. 8.4
Standard Error of Estimate (*Y*)	$S_{YX} = S_Y\sqrt{1 - r^2}$	Sec. 8.6
Binomial Expansion	$(P + Q)^N = P^N + \frac{N}{1}P^{N-1}\,Q + \frac{N(N-1)}{1(2)}P^{N-2}\,Q^2 + \frac{N(N-1)(N-2)}{1\ (2)(3)}P^{N-3}\,Q^3 + \cdots + Q^N$	Sec. 10.4

Standard Error of the Mean	$\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}}$	Sec. 11.4
General Formula for a *z* Score	$z = \dfrac{\text{score} - \text{mean}}{\text{standard deviation}}$	Sec. 12.5
Estimate of the Population Standard Deviation	$s_X = \sqrt{\dfrac{SS_X}{n-1}}$	Sec. 12.9
Estimate of the Standard Error of the Mean	$s_{\overline{X}} = \dfrac{s_X}{\sqrt{n}} = \dfrac{S_X}{\sqrt{n-1}}$	Sec. 12.9
t Test: Hypothesis about a Single Mean	$t = \dfrac{\overline{X} - \mu_{\text{hyp}}}{s_{\overline{X}}}$	Sec. 12.10
Effect Size	One sample: $d = \dfrac{\mu_{\text{true}} - \mu_{\text{hyp}}}{\sigma}$ or $g = \dfrac{\overline{X} - \mu_{\text{hyp}}}{s_X}$ Two samples: $d = \dfrac{(\mu_X - \mu_Y)_{\text{true}} - (\mu_X - \mu_Y)_{\text{hyp}}}{\sigma}$ or $g = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{\sqrt{\dfrac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}}}$ or $r = \sqrt{\dfrac{t^2}{t^2 + df}}$	Sec. 13.2 Sec. 14.8
Estimate of the Standard Error of the Difference between Two *Independent* Means	$s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2} = \sqrt{\dfrac{s_X^2}{n_X} + \dfrac{s_Y^2}{n_Y}}$	Sec. 14.4
Estimate of the Standard Error of the Difference between Two Independent Means (Assuming Homogeneity of Variance)	$s_{\overline{X}-\overline{Y}} = \sqrt{\dfrac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\left(\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right)}$	Sec. 14.4
(Same as above when $n_X = n_Y$)	$s_{\overline{X}-\overline{Y}} = \sqrt{\dfrac{SS_X + SS_Y}{n(n-1)}}$	Sec. 14.4
Test Statistic: Hypothesis about the Difference between Two *Independent* Means, *t* Distribution Model	$t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{\sqrt{\dfrac{SS_X + SS_Y}{(n_X - 1) + (n_Y - 1)}\left[\dfrac{1}{n_X} + \dfrac{1}{n_Y}\right]}}$	Sec. 14.4
Estimate of the Standard Error of the Difference between Two *Dependent* Sample Means	$s_{\overline{X}-\overline{Y}} = \sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2 - 2r s_{\overline{X}} s_{\overline{Y}}}$	Sec. 15.1
t Test: Hypothesis about the Difference between Two *Dependent* Means	$t = \dfrac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)_{\text{hyp}}}{\sqrt{s_{\overline{X}}^2 + s_{\overline{Y}}^2 - 2r s_{\overline{X}} s_{\overline{Y}}}}$	Sec. 15.1

(continued on back endpapers)

USEFUL FORMULAS

Here are some of the computing formulas that appear in the book. The reference is to the section number where each formula is introduced and defined.

Confidence Interval for μ_X	$\overline{X} \pm t_P s_{\overline{X}}$	Sec. 17.2
Confidence Interval for $\mu_X - \mu_Y$	$(\overline{X} - \overline{Y}) \pm t_P s_{\overline{X}-\overline{Y}}$	Sec. 17.7
Within-Groups Sum of Squares	$SS_{\text{w}} = \sum^{\text{all scores}} (X - \overline{X})^2$	Sec. 18.3
	$= \sum^{\text{all scores}} X^2 - \left[\frac{\left(\sum X_A\right)^2}{n_A} + \frac{\left(\sum X_B\right)^2}{n_B} + \cdots \right]$	Sec. 18.9
Between-Groups Sum of Squares	$SS_{\text{bet}} = \sum^{k} n_i (\overline{X}_i - \overline{\overline{X}})^2$	Sec. 18.3
	$= \left[\frac{\left(\sum X_A\right)^2}{n_A} + \frac{\left(\sum X_B\right)^2}{n_B} + \cdots \right] - \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}}$	Sec. 18.9
Total Sum of Squares	$SS_{\text{total}} = \sum^{\text{all scores}} (X - \overline{\overline{X}})^2$	Sec. 18.3
	$= \sum^{\text{all scores}} X^2 - \frac{\left(\sum^{\text{all scores}} X\right)^2}{n_{\text{total}}}$	Sec. 18.9
Within-Groups Degrees of Freedom	$df_{\text{w}} = n_{\text{total}} - k$	Sec. 18.4
Between-Groups Degrees of Freedom	$df_{\text{bet}} = k - 1$	Sec. 18.4
Within-Groups Variance Estimate	$s^2_{\text{w}} = \frac{SS_{\text{w}}}{df_{\text{w}}}$	Sec. 18.5
Between Groups Variance Estimate	$s^2_{\text{bet}} = \frac{SS_{\text{bet}}}{df_{\text{bet}}}$	Sec. 18.5
F Ratio	$F_{\text{calc}} = \frac{s^2_{\text{bet}}}{s^2_{\text{w}}}$	Sec. 18.5

Tukey's HSD Test	$\text{HSD} = q\sqrt{\dfrac{s_{\text{w}}^2}{n}}$	Sec. 18.14
Standard Error of a Planned Comparison	$s_k = \sqrt{s_{\text{w}}^2\left(\dfrac{a_A^2}{n_A} + \dfrac{a_B^2}{n_B} + \cdots + \dfrac{a_k^2}{n_k}\right)}$	Sec. 18.16
Column Sum of Squares	$SS_C = \sum^{\substack{\text{all}\\ \text{columns}}} n_C\,(\overline{X}_C - \overline{\overline{X}})^2$	Sec. 19.4
Row Sum of Squares	$SS_R = \sum^{\substack{\text{all}\\ \text{rows}}} n_R(\overline{X}_R - \overline{\overline{X}})^2$	Sec. 19.4
Within-Cells Sum of Squares	$SS_{\text{wc}} = \sum^{\substack{\text{all}\\ \text{scores}}} (X - \overline{X}_{\text{cell}})^2$	Sec. 19.4
$R \times C$ Interaction Sum of Squares	$SS_{R\times C} = SS_{\text{total}} - (SS_C + SS_R + SS_{\text{wc}})$	Sec. 19.4
Degrees of Freedom for Columns	$df_C = C - 1$	Sec. 19.5
Degrees of Freedom for Rows	$df_R = R - 1$	Sec. 19.5
Degrees of Freedom for $R \times C$ Interaction	$df_{R\times C} = (R - 1)(C - 1)$	Sec. 19.5
Within-Cells Degrees of Freedom	$df_{\text{wc}} = \sum^{\substack{\text{all}\\ \text{cells}}} (n_{\text{wc}} - 1)$	Sec. 19.5
Chi-Square: General Formula	$\chi^2 = \sum\left[\dfrac{(f_{\text{o}} - f_{\text{e}})^2}{f_{\text{e}}}\right]$	Sec. 20.2
Mann-Whitney *U* Test	$U_X = (n_X)(n_Y) + \dfrac{n_X(n_X + 1)}{2} - \sum R_X$	Sec. 21.5
Kruskal-Wallis Test	$\text{H} = -3(n_{\text{total}} + 1) + \dfrac{12}{n_{\text{total}}(n_{\text{total}} + 1)}$ $\left[\dfrac{\left(\sum R_1\right)^2}{n_1} + \dfrac{\left(\sum R_2\right)^2}{n_2} + \cdots \dfrac{\left(\sum R_k\right)^2}{n_k}\right]$	Sec. 21.8